



3 1761 10374367 0



Survey Methodology

Catalogue No. 12-001-XPB

A journal
published by
Statistics Canada

June 2006

•

Volume 32


•

Number 1



Statistics Canada
Statistique Canada

Canada



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743670>



Survey Methodology

A journal
published by
Statistics Canada

June 2006 • Volume 32 • Number 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2006

All rights reserved. This product cannot be
reproduced and/or transmitted to any person or organization outside of the licensee's organization.

Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes. This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows: Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s).

Otherwise, users shall seek prior written permission of Licensing Services, Client Services Division,
Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

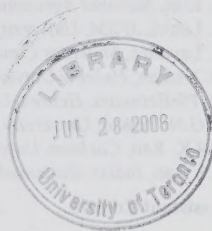
July 2006

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman D. Royce

Past Chairmen G.J. Brackstone
R. Platek

Members J. Gambino
R. Jones
J. Kovar
H. Mantel
E. Rancourt

EDITORIAL BOARD

Editor J. Kovar, *Statistics Canada*
Deputy Editor H. Mantel, *Statistics Canada*

Past Editor M.P. Singh

Associate Editors

D.A. Binder, *Statistics Canada*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidioglou, *Office for National Statistics*
D. Judkins, *Westat Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
G. Nathan, *Hebrew University*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*

J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
R. Valliant, *JPSM, University of Michigan*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, D. Haziza, Z. Patak, S. Rubin-Bleuer and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$30 (\$15 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada, the International Association for Official Statistics and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.ca.

Survey Methodology

A Journal Published by Statistics Canada

Volume 12, Number 1, June 2008

Contents

In This Issue	1
M. P. Singh: Remembered	2
Regular Papers	
Steven K. Thompson Doubtful Purposes, With Design	11
Samuel B. Thomas and Chris Shalizi Using Making Good Methods to Control for Measurement Error in a Double-Blind Experiment	25
Steven Harris and Peter D. Jacob On Callibratio	
Dedicated to the family of M.P. Singh: His wife, Savitri and his children, Mala, Mamta and Rahul	
David Harland and Jon V. K. Rao A Nonresponse Model Approach to Imputing Under-Reported or Missing Survey Data	37
John L. Zaretsky and Adam M. Zaretsky A Model for Forecasting and Improving Surveying Canadian Households under Sampling for Nonresponse Follow-up	65
Other Challenges	
The 2008 Vancouver Record Check Sample Attrition	77
Statistical Theory Toughest	
Sampling with Correlation in Small-Area Estimation	87
Long Vu and Thomas Chagnon Small Area Estimation Using Area Level Models and General Sampling Variates	99
All-Data Estimation and Multiple-Use Tables: Methods	
A Case-Deletion Strategy for Poststratification Estimation: A Small Area Approach	105
Other Notes	
Revised Optimal Sample Size and Power Loss Correction Factors for Multiple Design Samples	117

Survey Methodology
A journal Published by Statistics Canada
Volume 32, Number 1, June 2006

Contents

In This Issue.....	1
M.P. Singh Remembered.....	3

Regular Papers

Steven K. Thompson Targeted Random Walk Designs	11
Gabriele B. Durrant and Chris Skinner Using Missing Data Methods to Correct for Measurement Error in a Distribution Function.....	25
Torsten Harms and Pierre Duchesne On Calibration Estimation for Quantiles	37
David Haziza and Jon N.K. Rao A Nonresponse Model Approach to Inference Under Imputation for Missing Survey Data.....	53
Elaine L. Zanutto and Alan M. Zaslavsky A Model for Estimating and Imputing Nonrespondent Census Households under Sampling for Nonresponse Follow-up	65
Alain Th��berge The 2006 Reverse Record Check Sample Allocation.....	77
Nicholas Tibor Longford Sample Size Calculation for Small-Area Estimation.....	87
Yong You and Beatrice Chapman Small Area Estimation Using Area Level Models and Estimated Sampling Variances	97
Ali-Reza Khoshgooyanfar and Mohammad Taheri Monazzah A Cost-Effective Strategy for Provincial Unemployment Estimation: A Small Area Approach.....	105

Short Notes

Siegfried Gabler, Sabine H��der and Peter Lynn Design Effects for Multiple Design Samples	115
--	-----

In This Issue

This issue of the *Survey Methodology* journal opens with a special article to honour the memory of M.P. Singh, the founding Editor who led the journal for thirty years to its current stature as an internationally recognized source for new developments in survey methods and methods for production of official statistics. In this article many of M.P.'s closest colleagues and friends from over the years share their memories of him, and reflect on his career and his contributions.

In the first regular paper of this issue, Thompson discusses random walk designs for sampling from a networked population. He shows how this approach can lead to network samples where the inclusion probabilities can be estimated independently of how the initial sample of nodes is chosen, leading to valid design-based inference methods. Selection preference can be given to certain types of nodes or graph characteristics through choice of the random walk mechanism. He describes both uniform and targeted random walk designs, and presents some examples for illustration.

Durrant and Skinner consider the use of imputation and weighting to correct for measurement error in the estimation of a distribution function. They consider various nearest neighbor and hot-deck imputation methods, and propensity score weighting under different response models. They discuss the theoretical properties of these methods, and compare them via simulations to estimate the distribution of hourly pay based on United Kingdom Labour Force Survey data. They conclude that an approach based on fractional imputation seems best overall in terms of efficiency and robustness.

Harms and Duchesne look at the problem of estimation of quantiles using survey data. They calibrate an interpolated estimate of a distribution function to given quantiles of an auxiliary variable, and then invert the resulting calibrated interpolated estimator of the distribution function of the variable of interest. They compare their approach with other methods in a simulation study.

In their paper, Haziza and Rao propose a new regression imputation method that uses the response probabilities. The new method leads to valid estimators under either the nonresponse model approach or the imputation model approach. In the nonresponse model approach, the response mechanism is parametrically modelled and is not restricted to the uniform nonresponse model, while in the imputation model approach the variables of interest are modelled and nonresponse is assumed to be ignorable. The authors also provide estimators of the variance under their imputation method. Simulation results for both point and variance estimation are reported that show the good performance of the proposed regression imputation method.

The paper by Zanutto and Zaslavsky deals with the problem of estimation in the U.S. decennial census of population under sampling for nonresponse follow-up. Instead of trying to obtain information from all nonrespondents, a sample is drawn for follow-up, thus creating a small area estimation problem. The proposed strategy consists of predicting the number of nonrespondent households in different categories using a hierarchical loglinear model and then imputing detailed person and household information using donor imputation. The idea in the first step is to model household characteristics using low-dimensional covariates at detailed levels of geography and more detailed covariates at higher levels of geography. The performance of the proposed model compares favourably to other models in a simulation study.

In Théberge's article, a new approach is proposed for 2006 Reverse Record Check (RRC) sample allocation for measuring census undercoverage and a part of overcoverage. RRC estimates are used together with census counts to produce population estimates which will be used in calculating Canadian federal government equalization payments to the provinces. The proposed approach will provide an allocation that achieves the proper balance between four objectives. It first consists of establishing a separate allocation for each objective. Then, by province, the maximum sample size is used for each allocation. Finally, the RRC's sub-provincial sample allocation is obtained by using calibration to smooth the stratum-level parameters.

In his paper Longford discusses how to design a survey when estimates are required for a number of small areas, with possibly different priorities for different small areas, by minimizing a weighted sum of expected variances. He first develops his ideas for direct estimation, and then extends to composite estimation that combines the direct estimator with a synthetic estimator. The approach is illustrated by the resulting sample allocations, under various assumptions, to cantons in a Swiss household survey.

You and Chapman propose a Hierarchical Bayes estimation approach for small area estimation when the sampling errors of direct estimators are estimated. They demonstrate the approach by producing small area estimates from two data sets and investigate the sensitivity of their approach to model assumptions.

Khoshgooyanfar and Monazzah compare synthetic, composite and Empirical Bayes small area estimation methods for producing intercensal estimates of unemployment rates for provinces in Iran. They find that both composite and Empirical Bayes approaches lead to satisfactory results.

The short note by Gabler, Häder and Lynn, the last paper in this issue, provides an interesting extension to the earlier paper by Gabler, Häder and Lahiri that appeared in *Survey Methodology* (1999). It offers a practical solution for obtaining design effects when different exclusive domains use different sample problems.

Finally, we note that *Survey Methodology* is now available on-line in a fully searchable pdf format. All articles published in the journal are now being made available free of charge directly on the Statistics Canada web site upon release. There are also plans to include past issues. All the articles from the latest seven issues have been posted, and work is in progress to add those of the previous 10 years. Printed copies of the journal will still be produced for subscribers. Older issues can be obtained upon request in paper or pdf scanned formats. The journal can be accessed from Statistics Canada's web site at www.statcan.ca/bsolc/english/bsolc?catno=12-001-X.

Harold Mantel, Deputy Editor

M.P. Singh Remembered

Introduction

Don Royce
Statistics Canada

In August of 2005 the world of survey methodology lost one of its leading figures with the death of Dr. M.P. Singh at the age of sixty-three, just a few months short of his planned retirement. M.P. and I had discussed his upcoming retirement only briefly, but it was intuitively clear to both of us that he would continue as the Editor of *Survey Methodology* even after he left Statistics Canada. *Survey Methodology* was a part of his life, and I was only too happy to offer M.P. the chance to work part-time from his family's home in Toronto so that he could continue to nurture the journal that he had led for over thirty years. Sadly, this arrangement never came to be realized.

In the series of articles that follow, many of M.P.'s closest colleagues and friends (the two are indistinguishable) recall M.P. Singh the statistician, editor, collaborator, leader, and human being. I am deeply indebted to Eric Rancourt of Statistics Canada for suggesting this series of articles, and to all of the authors who gave their time and talents to put into words their memories of M.P. Singh. Although words can never completely capture the essence of a person, the articles that follow do a marvellous job of describing the life of M.P. Singh and remind us of the legacy he left to all of us who were fortunate enough to know him. We hope M.P. would be pleased.

Some Reminiscences

J.N.K. Rao
Carleton University, Ottawa

I first met Mangala Prasad Singh (fondly known to many of us as M.P.) in 1968 while I was a visiting professor at the Indian Statistical Institute (ISI), Calcutta. M.P. was a Ph.D. student at the ISI working under the supervision of M.N. Murthy. While doing his Ph.D. he also worked in the National Sample Survey (NSS) of India. NSS was located in the ISI campus and M.P. worked under renowned survey statisticians at the NSS and ISI including P.C. Mahalanobis, D.B. Lahiri and M.N. Murthy. He received solid training in both design and theory of sample surveys. M.P. made good use of that sound training throughout his illustrious career

by following the principles of efficient design subject to cost and operational considerations and insisting on sound theory before implementing new survey designs or redesigning continuing surveys such as the Canadian Labour Force Survey (LFS).

A major part of M.P. Singh's thesis was on the efficient use of auxiliary information. He studied the case of two auxiliary variables, one positively correlated and the other negatively correlated with the variable of interest, and developed ratio-cum-product estimators of totals. Murthy (1967) devoted a section in his well-known sampling book to ratio-cum-product estimators. M.P. published several papers on the efficient use of auxiliary information based on his thesis work: ratio-cum-product estimators (*Metrika* 1967; *Sankhyā* 1969), multivariate product estimation (*Journal of the Indian Society of Agricultural Statistics* 1967) and systematic sampling in ratio and product estimation (*Metrika* 1967). He also published an important paper in the *Annals of Statistics*, 1967 on the relative efficiency of two-phase sampling strategies under a super-population model. The first phase consisted of simple random sampling to collect data on an auxiliary variable x that was used at the second phase to select a PPS sample without replacement and collect data on the variable of interest, y .

M.P. was also dabbling with inferential issues in survey sampling at the time of my visit to ISI and he encountered technical problems in proving some admissibility results: An estimator is admissible in a class of unbiased estimators if no other estimator in the class is uniformly more efficient. Unfortunately, the criterion of admissibility is not sufficiently selective and as a result other admissibility related criteria for unique choice were proposed in the literature. I was also interested in inferential issues at that time and this led to our collaboration on admissibility related topics. The resulting work constituted a part of his Ph.D. thesis. We ultimately published a paper based on this work in the *Australian Journal of Statistics*, 1973 based on our 1969 ISI Technical Report. Our results demonstrated the practical irrelevance of a criterion called hyper-admissibility that leads to the Horvitz-Thompson (HT) estimator of total as the *unique* choice for *any* sampling design. Subsequently, D. Basu obtained similar results independently in his 1971 landmark paper on inferential issues, and his famous example of circus elephants put a stop to research on unrealistic criteria that lead to unique choice for any design.

M.P. also showed that hyper-admissibility when applied to variance estimation leads to a “bad” variance estimator as the unique choice.

Soon after joining Statistics Canada in 1970 as a Methodologist, M.P. was actively involved in the redesign of the LFS that led to several innovations. M.P. proposed the use of systematic PPS sampling without replacement with initial randomization for selecting the primary units from the non-self-representing units (NSRUs) and the random group method with one primary unit from each random group selected by PPS sampling from the self-representing units (SRUs). In the 1960s I studied those methods theoretically from the point of view of efficiency and variance estimation. M.P. on the other hand recognized their practical advantages in the context of LFS. Both systematic PPS sampling and random group method permitted sample expansion as well as easy rotation of sample primary units over time and the random group method enabled the adaptation of Keyfitz’ ingenious method for changing out-dated size measures within each random group. A joint paper with Dick Platek on updating size measures was published in *Metrika*, 1975. The LFS group under the able guidance of M.P. made several methodological advances to improve the efficiency of the design as well as estimation. Given M.P. Singh’s past interest in the effective use of auxiliary information, the LFS switched to generalized regression estimation to accommodate several post-stratification variables. The LFS group also was the first to recognize the merits of re-sampling variance estimation and the jackknife was adopted for variance estimation. More recently, regression composite estimation was introduced in the LFS under M.P. Singh’s leadership, using a method suggested by Wayne Fuller and myself that is good for both change and level estimation. This method and an earlier method of Avi Singh fit in well with the existing LFS estimation system based on generalized regression. Three papers on regression composite estimation for LFS, including a joint paper of M.P. with Jack Gambino and Brian Kennedy, were published in the June 2001 issue of *Survey Methodology*.

M.P. also had a keen interest in small area estimation, dating back to 1976. His team made important methodological contributions to small area estimation. M.P. and his colleagues proposed simple synthetic estimators as well as a new estimator called the sample dependent estimator. The latter estimator is a simple composite estimator with weights designed to account for realized sample sizes smaller than expected sample sizes in the areas. Sample dependent estimators became quite popular and many agencies worldwide have used them. M.P. Singh’s 1994 joint paper in *Survey Methodology* with Jack Gambino and Harold Mantel addresses several practical issues pertaining to small area

estimation. I particularly like the section on design issues. It presents an excellent illustration of compromise sample allocation in the LFS to satisfy reliability requirements at the provincial level as well as sub-provincial level. A section in my 2003 Wiley book on Small Area Estimation is devoted to design issues largely based on the 1994 paper. M.P. played an active role in organizing a highly successful international conference on Small Area Estimation in 1985 and acted as co-editor of a 1987 book *Small Area Statistics* published by Wiley based on the invited papers presented at the conference.

M.P. thoroughly enjoyed working as Editor-in-Chief of *Survey Methodology*. He maintained close contact with his team of Associate Editors and introduced many innovative ideas including theme papers on both theory and practice and the Waksberg series of papers. The luncheon gatherings M.P. organized at the Annual Joint Statistical Meetings were always a big hit with the Associate Editors! As an Associate Editor located in Ottawa and consultant to Statistics Canada, I had many conversations with M.P. on matters related to the journal over the past 25 years. M.P. also played an active role in the Statistical Society of Canada (SSC) and he was instrumental in raising the profile of survey sampling at the SSC Annual Meetings.

M.P. was remarkably accurate in palm reading. In 1999 he read my palms and warned me of health problems. Indeed, I faced an unexpected health problem in 2001 due to complications from appendicitis. A few months before M.P.’s death, Avi Singh told me that M.P. had read his own palms and predicted recovery from his serious health problems. Both Avi and I were very confident that we would see M.P. back at work. However, it is a common belief in India that palmists reading their own hands cannot predict their futures accurately. Unfortunately, this belief proved to be true in this instance.

M.P. was truly a great friend of mine and I will miss him very much. It is fitting that his ashes were immersed in the sacred river Ganges in the holiest city for Hindus, Varanasi (also called Benares), where M.P. was born. His soul has gone to Heaven but his legacy will remain with us.

M.P. and his Research Days

T.J. Rao

Indian Statistical Institute, Kolkata

I had first met M.P. when he came to attend the Fourth Summer Course (Advanced) for Statisticians organized by the Research and Training School (RTS) of the Indian Statistical Institute (ISI) in May–June 1964 at the University of Kerala in the South Indian city of Trivandrum (now Thiruvananthapuram). This course was meant for research

scholars and junior faculty of ISI and other Universities. M.P. came from the Benares Hindu University (BHU) where he was a temporary lecturer. He obtained his Bachelor's degree in Statistics from the same University (BHU) and a Masters from University of Poona. I was among the research scholars that were selected from ISI for this course. We did not have much interaction during the course.

A little later, M.P. was offered a job in the Sampling Division of the National Sample Survey (NSS) Department, which at that time was part of ISI. Professors D.B. Lahiri, S. Rajarao and M.N. Murthy among others were already heading several divisions of NSS by then. Besides being occupied with the designing of the large scale sample surveys conducted by NSS, M.P. spent his spare time on research problems in sample surveys. Lahiri and Murthy encouraged methodological research in the NSS and had started a seminar series as well as release of technical reports similar to the RTS technical reports of ISI. M.P. and I spoke on our research on sampling problems in these seminars organized by NSS as well as RTS. Most of the work of M.P., which he made into technical reports of the NSS Series, got published later on in well known journals.

With his expertise in the NSS on multi purpose surveys, he got interested in the problems of utilization of auxiliary information in sample surveys. His early work related to ratio and product methods of estimation. M.P. successfully and intelligently considered the case of multiple auxiliary variables of which some are positively correlated and some negatively correlated with the study variable and used ratio estimators for the former and product estimators for the latter and produced the "ratio cum product estimator" (Singh 1967). This paper is often quoted and several scholars, especially from India, published extensions. Jointly with M.N. Murthy, he developed interesting concepts of admissibility of estimators (Murthy and Singh 1969). During the year 1968, Professor J.N.K. Rao visited ISI and we were very fortunate to have interaction with him.

M.P. was very much interested in attending conferences. He never missed any at his alma mater BHU nor the sessions of the Indian Science Congress. He took the task of writing his thesis very seriously and used to have discussions with Professors M.N. Murthy, J.N.K. Rao and D. Basu. He submitted his research work as a Thesis (Singh 1969) for the degree of Doctor of Philosophy (Ph.D.) of the Indian Statistical Institute in 1969 under the general guidance of M.N. Murthy. He left the NSS and ISI in 1970 to join Statistics Canada.

All the research scholars of ISI during 1965–70 and his colleagues at the NSS miss him very much.

References

- Singh, M.P. (1967). Ratio cum Product Method of Estimation. *Metrika*, 12, 34–42.
- Singh, M.P. (1969). *Some aspects of Estimation in Sampling from Finite Populations*. Ph.D. Thesis submitted to the Indian Statistical Institute.
- Murthy, M.N., and Singh, M.P. (1969). On the Concepts of Best and Admissible Estimators in Sampling Theory. *Sankhyā*, 31, 343–354.

M.P. Singh

Nanjamma Chinnappa Statistics Canada (retired)

While many know M.P. the statistician and of his achievements in statistics, I will try to write about M.P. the man.

I had not met M.P. until I came to Canada, although I had heard that he was the young man appointed in my position when I resigned my job at the National Sample Survey (NSS) department of the Indian Statistical Institute in Kolkata, India. I heard that when Dr. M.N. Murthy (then the head of the Methodology area in the NSS) sent me the draft of his book *Sampling Theory and Methods* for review, M.P. was the one who read my comments and discussed them with Dr. Murthy. Much later, when Dr. Murthy heard that I was hired by Statistics Canada, he gave me M.P.'s telephone number in Ottawa. So, when we arrived in Ottawa, I called M.P. from the hotel we were put up in and to my surprise he drove to the hotel on a cold, damp morning in late September and took me to Statistics Canada. That warm and friendly gesture brightened my day and my introduction to Statistics Canada.

M.P. hailed from the ancient city of Benares in India and it would appear that some of the qualities for which that city is famous had rubbed off on him. He was gentle, friendly to all, unflappable, resilient and wise. Many have told me how he was never too busy to listen to their problems and always helped with kind words and suggestions. Many young statisticians have benefited from his advice related to their research and career.

M.P. was fond of classical Indian music and dance. A family-oriented man, he was a pillar of strength for his wife and children during their times of need. At social gatherings he was full of fun and laughter. And when he first fell seriously ill some years ago he told me that it was his faith in God and in himself that helped him to recover. He will long be remembered, not only as a statistician of repute but as a good man who befriended and helped many.

A Career in Survey Methodology

Gordon Brackstone

Statistics Canada (retired)

M.P. Singh spent almost his whole career in the methodology area of Statistics Canada. He joined the organization in 1970, after obtaining a Ph.D. in survey sampling from the Indian Statistical Institute. At the time of his death he was Director of the Household Survey Methods Division in the Methodology Branch. His rise through the organization was steady rather than meteoric: he became a section Chief in 1973, an Assistant Director in 1982, and a Director in 1994. This steady progression mirrored his approach to survey methodology which valued thoroughness in research and testing to build firm foundations for implementation and further improvement.

Our careers at Statistics Canada coincided, give or take a year at either end, and intersected frequently, particularly from 1982 onwards. In the early 1980s when we felt the need to improve integration and oversight of Statistics Canada's methodology research work, there was little doubt in my mind who we would ask to head this effort and M.P. was duly appointed as the first Chair of the Methodology Research Committee. In this role until 1987 he initiated the planning processes and reporting requirements that, with further improvements from his successors, have governed the management of methodology research for two decades. It was during this same period that Statistics Canada's annual methodology symposia became established, with M.P. playing a key role in several of the earliest symposia (and many more subsequently).

In his long career at Statistics Canada, M.P. was involved in a broad range of methodological work, but his name will always be most closely associated with two projects: the design of the Canadian Labour Force Survey (LFS), and the Editorship of the journal, *Survey Methodology*.

The LFS provides the foundation for Statistics Canada's household survey program. Not only is it the source of monthly estimates of labour market conditions in Canada, but its frame is also the sampling basis for many other household surveys, including several longitudinal surveys introduced in the 1990s. Its efficient design is therefore crucial to the cost-effectiveness of Canada's social statistics program. First introduced in 1945, the LFS has typically undergone at least a sample redesign after each decennial Census. M.P. happened to join Statistics Canada just in time for the major post-1971 Census redesign. This redesign encompassed not only the sampling scheme, but also the questionnaire, the methods of collection, and the processing systems. Such a major redesign required extensive interdisciplinary project teamwork and M.P. became a key

player in the methodological aspects of this redesign. His papers from that period focus on optimizing the multi-stage design and updating the sample. He was co-author of the official description of the methodology of the Canadian Labour Force Survey (Platek and Singh 1976).

Following this redesign pressure to produce labour market estimates for smaller regions increased. This led him to develop methods for small area estimation from the LFS (Drew, Singh and Choudhry 1982). By the time of the post-1981 Census redesign, M.P. had become the Chair of the Redesign Committee responsible for oversight of the whole redesign. In addition to the usual sampling efficiency objectives, this redesign aimed to produce better sub-provincial data and to enhance the role of the LFS as a vehicle for conducting other household surveys. Naturally M.P. was again a principal author of the description of the new design (Statistics Canada 1990).

The efforts to make the LFS frame a basis for other household surveys were so successful that by the late 1990s a problem of overload had arisen. With the introduction of longitudinal surveys in addition to the regular survey program, concerns over the burden on the frame were increasing. In addition, the need for more targeted survey frames for certain sub-populations was being felt. M.P., set about finding alternative approaches, including approaches that would take advantage of the address register being developed for Census purposes. Some of these approaches were incorporated into the post-2001 redesign of the LFS that was just being introduced at the time of his death; some more ambitious ideas for a new frame for household surveys are still under consideration by his successors.

For more than 30 years M.P. guided methodological input to the LFS. His many papers, often co-authored with his staff, bear witness to his lasting imprint on the design of this flagship survey, and his guidance of many younger statisticians in the early stages of their careers.

Over this same period, M.P. also bore another heavy responsibility as Editor of *Survey Methodology*. The evolution of this journal from its inception in 1975 to its 25th Anniversary has been described by its founder, Richard Platek (1999), who had the foresight to appoint M.P. as its first Editor.

Under M.P.'s leadership the journal passed many milestones. In 1982 it became an official Statistics Canada publication – fully bilingual and priced. Authorship was expanded beyond Statistics Canada employees; a highly qualified panel of associate editors was recruited; theme issues were introduced, often attracting the best papers from a recent conference or symposium; the Editor's *In This Issue* feature was introduced to provide an overview of content; special 25th anniversary issues were published in 1999 – 2000, along with an index for Volumes 1–26. Over

this period, arrangements were negotiated, firstly with the International Association of Survey Statisticians and later with other statistical societies, to provide discounted subscriptions. More recently electronic versions of the journal have been made available.

Throughout these developments M.P. was at the helm, planning future issues, on guard for interesting research worthy of inclusion, encouraging potential authors, recruiting and pestering associate editors through the refereeing process, working with Statistics Canada's publication and marketing staffs to improve and promote the journal. On the journal's Management Board from 1987-2004, I witnessed first-hand and admired his enthusiasm and perseverance in the face of many difficulties. It was for him, I believe, a true labour of love.

These brief descriptions of just two of M.P.'s many contributions to Statistics Canada and the statistics profession cannot do full justice to his career. I hope they give an impression of an ever dependable professional who combined a deep understanding and research ability in statistical methods with an appreciation of the practical constraints of applying statistical methods to surveys. His style was based on reason and persistence, without bluster and shunning confrontation, coupled with an innate concern for the feelings of others. It was always a pleasure to work with M.P. and an honour to be associated with his accomplishments.

References

- Drew, D., Singh, M.P. and Choudhry, H. (1982). Evaluation of Small Area Estimation Techniques. *Survey Methodology*, 8, 17-47.
- Platek, R., and Singh, M.P. (1976). *Methodology of the Canadian Labour Force Survey*, Statistics Canada, Catalogue number 71-526.
- Platek, R. (1999). Survey Methodology – The First 25 Years. *Survey Methodology*, 25, 109-111.
- Statistics Canada (1990). *Methodology of the Canadian Labour Force Survey 1984-1990*, Statistics Canada, Catalogue number 71-526.

In Memory of M.P. Singh

Fritz Scheuren

2005 President, American Statistical Association

In M.P. Singh last summer we lost an individual known throughout the whole statistical world as a scholar, a gentleman, and a doer. When I spoke about him at the fall 2005 Statistics Canada Methodology Symposium, it was from this perspective.

I will be brief however, providing only a sample of what could be said. Others are writing too. I will leave it to them to say more.

My memories of M.P. go back over 20 years. Exactly when we first met is now obscure to me but I have been one of his associate editors (AE's) at *Survey Methodology* for at least that long.

He used to like to have me look at papers on record linkage, sometimes sample weighting or estimation, and, less commonly, on missing data topics. His selections were ones I invariably learned from. By and large, after his initial screening, the incoming quality was excellent and, working under him my job was to make sure that the journal versions that eventually resulted were even better.

His editorship of *Survey Methodology* was challenging. The Journal had to have closely argued mathematical statistical formulations but these also had to be ones that could be put into practice. In other words the ideas had to be very good, as well as eminently useful. And they have consistently been both. No mean feat.

Many outstanding younger professionals, when they first submit a paper, demonstrate just one of these two attributes in their submissions, usually the mathematical side of their topic. For submissions that achieved at least one of these, my interpretation of the goal M.P. set for his AE's was to help authors, through the referee and AE comments, to achieve the second goal too. And what a journal he created with his vision!

By the way, he suggested that I might have tended to overdo my author support role but I think that secretly he was pleased with my approach of never giving up on what could become a great paper, if given patience. And there were several papers I handled that his patience was tried but eventually rewarded in the end.

M.P. had toughness, though, that complemented his unfailing gentleness. He firmly held all of us to high standards in guiding *Survey Methodology* with a sure hand. Even when his health began to fail, his spirit always remained visible.

The one word summary I used to characterize M.P. at the fall conference was to call him a "Mensch." Now this German word for "person" may be familiar to many of you in its Yiddish sense of a complete or whole human being. But frankly "Mensch" is really untranslatable. That is why it has stayed in Yiddish here (although I have not written in Hebrew letters, as would have been appropriate). Certainly no simple definition can do justice to either the word or the individual that M.P. was.

We all miss him greatly. He was a good friend, a loving family man, open to new ideas, careful in his advice about practice and rigorous in his thinking. M.P. will forever be a model of what it means to be a sampling statistician.

Some Recollections of M.P. Singh

David A. Binder

Statistics Canada (retired)

My memories of M.P. Singh over the many years that I knew him are all very fond. His strengths, both as an outstanding survey statistician, and as a kind and gentle person, were characteristics that were unmatched.

It was in the summer of 1970 when I first met M.P. Singh. I was working as a summer student in Agriculture Division at Statistics Canada. M.P. Singh and J.C. (John) Koop were the methodologists working with Agriculture Division at the time. I was sharing an office with Jack Graham who was on sabbatical leave from Carleton University. Jack's comment to me at that time was how fortunate Statistics Canada was to have M.P. and John there as survey methodologists, as they were two of the finest survey statisticians in the world. In fact, it was such outstanding talent at Statistics Canada that helped me decide that it would be a good place to start my career.

Most people knew M.P. through his dozens of published papers, his stewardship of the journal, *Survey Methodology*, and his interventions at statistical conferences. His publications included papers on household survey designs and redesigns, estimation (including composite estimation and domain estimation), small area estimation, and nonresponse adjustment. His insights into the many complexities of survey methods were often reflected by his questions and suggestions at conferences and meetings.

He also co-edited monographs on panel surveys (Kasprzyk *et al.* 1989) and on small area statistics (Platek *et al.* 1987), and he wrote a review article on *Survey Methodology* in the *Encyclopedia of Statistical Sciences* (Singh 1988).

As editor of *Survey Methodology* since its inception in 1975, M.P. oversaw the evolution of the journal from its beginnings as mainly a vehicle for staff at Statistics Canada to publish their research to a top international journal with regular contributions from around the world. *Survey Methodology* has been adopted by the Section on Survey Research Methods of the American Statistical Association, and by the International Association of Survey Statisticians as a publication for members of those organizations. This is a reflection of the many years of M.P.'s "labour of love" on the journal. His gentleness and kindness were even reflected in his encouraging remarks when writing a letter of rejection to authors!

Over the years M.P. was a leader in adapting to the changing technology for households surveys. He always pursued ways to improve data collection methods. He guided Statistics Canada through the world of face-face interviewing, into telephone interviewing, and computer-assisted methods.

Most recently, he was keen to develop methods to improve efficiency by introducing the concept of a master sample for household survey designs at Statistics Canada, and he was instrumental in convincing managers from across Statistics Canada of the potential merits of this concept.

M.P. was a major influence at Statistics Canada to ensure the quality and the stature of research in Statistical Methods. The Bureau's accomplishments in this area have received recognition from around the world, and Statistics Canada is now often asked to participate in research activities, such as presenting invited papers at meetings, participating on panel discussions, and joining various advisory committees and panels. To help achieve this stature, the Methodology Research Committee was created in 1982-1983, with M.P. as its first chair. There he helped develop a research agenda and a strategic plan for the Methodology Branch. Although the research agenda has changed over time, the Methodology Research Program is still flourishing, thanks to the management structure and support that M.P. helped put in place.

Throughout my career at Statistics Canada, I was able to benefit greatly from M.P.'s presence. At management meetings and at meetings where he represented Methodology management, he always ensured that we kept our distinctiveness as methodologists, ensuring that decisions we took made sense for our group.

Even with all of M.P.'s accomplishments as a survey statistician, it was his character that I admired the most. His selfless compassion for others, no matter what their level of competence, was his greatest strength, in my opinion. I can recall one occasion when the two of us were interviewing a highly qualified candidate whom we brought to Ottawa from a fair distance away. However, after just a few minutes, it was clear that, in spite of this person's qualifications, he was not suitable for a position in the Methodology Branch. Yet, M.P. managed to make the candidate feel comfortable, after having made a special trip to Ottawa for the interview, by discussing that which the candidate was most familiar with, even though M.P. also recognized that the candidate was unsuitable for the Branch.

M.P. always praised others when their accomplishments were noteworthy. This is one of the many reasons why he was endeared to so many, and why so many will miss him.

References

- Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M.P. (Ed.) (1989). *Panel Surveys*. New York: John Wiley & Sons, Inc.
- Platek, R., Rao, J.N.K., Särndal, C.-E. and Singh, M.P. (Ed.) (1987). *Small Area Statistics: An International Symposium*. New York: John Wiley & Sons, Inc.
- Singh, M.P. (1988). *Encyclopedia of Statistical Sciences*, Vol. 9, (Eds. D.L. Banks, Read, B. Campbell and S. Kotz), 109-110. New York: John Wiley & Sons, Inc.

Manager and Mentor

Jack Gambino
Statistics Canada

Others have written of M.P. Singh's important and varied contributions to the statistics profession and to Statistics Canada. I had the good fortune to work closely with M.P. for 17 years and got to know a side of him that only those who worked with him on a regular basis saw and appreciated. I saw M.P. in his role as editor of *Survey Methodology*, including his involvement in the day-to-day activities that led to each issue of the journal, in his role as manager, and in his role as supervisor and mentor.

In the 1980s, when I first joined Statistics Canada, it was impossible not to come across M.P. Singh. To me, for the first few years, he was the person who asked probing questions at each and every methodology seminar I attended. Much later, when we happened to sit on some of the same committees, I was always amazed when, during meetings, he would come up with good questions on topics that were clearly not on methodology turf. Invariably, his questions helped to clarify the issues, not only for methodologists, but for everyone in attendance. The lesson I learned from this was not to assume that I'm the only one who doesn't fully understand the topic under discussion.

M.P. the Editor: I first got to know M.P. personally when I joined his subdivision in 1988. He immediately recruited me as an assistant editor of *Survey Methodology*. This was standard practice for M.P. – when people with a strong technical background came into his sights, they became potential assistant editors for the journal. Those of us lucky enough to become assistant editors learned a great deal from the experience. As M.P. grew to trust our judgment over time, he relied increasingly on our views, for example, in dealing with a paper that had received conflicting referee reports.

M.P. the Manager: M.P.'s approach to assistant editors is illustrative of how he managed more generally. He let people prove themselves and, with rare exceptions, each employee's abilities grew in parallel with M.P.'s confidence in him or her. Many managers follow a specific management philosophy, sometimes jumping on whatever the latest management fad is. M.P. was not in that category. He was an intuitive manager and had a knack for spotting future "talent" early in their careers. He was also a non-authoritarian manager who encouraged his staff in their work. Although M.P. was an open, easygoing manager, he knew when to put his foot down, as many of us who worked with him found out the hard way, albeit on rare occasions.

M.P. was a strategic thinker who liked to discuss both statistical and management issues thoroughly. This

sometimes led to long meetings where we were all expected to give our views. And just when we would think that an issue was settled, M.P. would throw in a new twist that got the discussion going again! The advantage of M.P.'s approach, of course, was that by the end of the meeting we all understood the ins and outs of the subject under discussion and almost always reached a consensus.

Throughout his career, M.P. took a strong interest in the development of researchers and the research function at Statistics Canada. He viewed an active research program as essential for the continued success of Statistics Canada. As a result, he worked to increase the professional visibility of researchers, and more generally of survey methodologists, within the Statistical Society of Canada and other organizations.

M.P. the supervisor and mentor: After working in M.P.'s area for a few years, I had the good fortune to report directly to him. Separating M.P. the supervisor from M.P. the mentor is impossible. He took a keen interest in his immediate employees' careers, giving them advice and steering them toward the right choices or, more importantly, steering them away from the wrong ones. What was interesting was the way he often did this. Rather than be direct, he would often lead the employee, in a near-Socratic way, to the realization that something was not such a good idea. Another technique was the "look" – anyone who got to know M.P. well learned to tell at a glance when M.P. thought an idea was particularly bad.

I learned a great deal about surveys from M.P. but more importantly, I think, I learned from him what makes a good manager, motivator and mentor. Thus I come to the realization that perhaps his greatest role was *M.P. the teacher*. Those of us who worked closely with M.P. over the years will continue to benefit from his example for the rest of our careers, and I expect we will pass on what we learned from him, filtered through our own unique experiences, to the next generation as well.

In his Own Words

Eric Rancourt
Statistics Canada

M.P. was a man of impressive personality. Many of his employees and colleagues did not have the chance to work closely with him, but for those who did, M.P. would reveal himself as a very comprehensive and human character. Below are a few quotes from him that others and I have collected. These words usually came to us at a comforting time and always made us come out of his office on a positive note.

- Don't bother setting up a meeting, my door is always open to discuss anything.
 - It's good to have a pet project.
 - We don't design surveys to calculate the variance.
 - I'm sure it can be done.
 - You're telling me that 2 out of 3 of your findings did not make it to the survey! Don't complain; if as much as 10% of your ideas get implemented, you'll have a great career!
- There is a sign by the highway that says 100 km/h; that doesn't mean you have to go to 100 km/h.
 - Don't worry, there is still time.
 - After all the efforts we make in designing surveys, what we remember and appreciate the most is not the methods or results; it is the people we worked with.

Targeted Random Walk Designs

Steven K. Thompson¹

Abstract

Hidden human populations, the Internet, and other networked structures conceptualized mathematically as graphs are inherently hard to sample by conventional means, and the most effective study designs usually involve procedures that select the sample by adaptively following links from one node to another. Sample data obtained in such studies are generally not representative at face value of the larger population of interest. However, a number of design and model based methods are now available for effective inference from such samples. The design based methods have the advantage that they do not depend on an assumed population model, but do depend for their validity on the design being implemented in a controlled and known way, which can be difficult or impossible in practice. The model based methods allow greater flexibility in the design, but depend on modeling of the population using stochastic graph models and also depend on the design being ignorable or of known form so that it can be included in the likelihood or Bayes equations. For both the design and the model based methods, the weak point often is the lack of control in how the initial sample is obtained, from which link-tracing commences. The designs described in this paper offer a third way, in which the sample selection probabilities become step by step less dependent on the initial sample selection. A Markov chain "random walk" model idealizes the natural design tendencies of a link-tracing selection sequence through a graph. This paper introduces uniform and targeted walk designs in which the random walk is nudged at each step to produce a design with the desired stationary probabilities. A sample is thus obtained that in important respects is representative at face value of the larger population of interest, or that requires only simple weighting factors to make it so.

Key Words: Adaptive sampling; Link-tracing designs; Markov chain Monte Carlo; Network sampling; Random walk; Respondent-driven sampling; Sampling in graphs; Sampling hidden population.

1. Introduction

Populations with linkage or network structure are conceptualized as graphs, with the nodes of the graph representing the units of the population and the edges or arcs of the graph representing the relationships or links between the units in the population. A central problem of studies in graph settings is that for many of the populations of interest it is difficult or impossible to obtain samples using conventional designs, and the samples obtained may be at face value highly unrepresentative of the larger population of interest. In practice, often the only practical methods of obtaining the sample involve following links from sample nodes to add more nodes and links to the sample. For example, in studies of hidden human populations such as injection drug users, sex workers, and others at risk for HIV/AIDS or hepatitis C, social links are followed from initially identified respondents to add more research participants to the sample. Similarly, in investigations of the characteristics of the Internet, the usual procedure is to obtain a sample of web sites by following links from initial sites to other sites.

Klov Dahl (1989) used the term "random walk" to describe a procedure for obtaining a sample from a hidden population by asking a respondent to identify several contacts, one of whom is selected at random to be the next respondent, with the pattern continuing for a number of steps. Heckathorn

(1997) described methods of "respondent-driven sampling" using procedures of this type. The motivation for using designs like this in practice is to penetrate deeper into the hidden population to obtain respondents who are more "representative" of the population than the more conspicuous initial respondents may be. In studies of the Internet, the parallel idea is that of the "random surfer", who selects a web page at random, clicks at random on one of the links on that page, thus moving to another page, and so on (Brin and Page 1998). The random walk design can be conceptualized as a Markov chain (Heckathorn 1997, 2002, Henzinger, Heydon, Mitzenmacher and Najork 2000, Salganik and Heckathorn 2004). In this paper some modifications of these Markov chain designs are described, with the object of obtaining stationary probabilities of equal or specified values in order to obtain simple estimates of characteristics of the population graph of interest.

Approaches to inference from samples in a graph setting include design-based, model-based, and combination methods. In the design based approach, all values of node and link variables in the graph are considered fixed or given, and inference is based on the design-induced probabilities involved in selecting the sample. In the model based approach, the population is itself viewed as a realization of a stochastic graph model, which provides the joint probability distribution of all the node and link variables. Previous design-based approaches include the methods of network or

1. Steven K. Thompson, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada, V5A 1S6. E-mail: thompson@stat.sfu.ca.

multiplicity sampling (Birnbaum and Sirken 1965), adaptive cluster sampling applied in a graph setting (Thompson and Collins 2002), and a few of the methods in the snowball sampling literature (Frank 1977, 1978, Frank and Snijders 1994). A method combining design and model based approaches is used in Felix-Medina and Thompson (2004) for studying a hidden population in which link-tracing follows from a probability survey sample from a frame that covers only part of the population.

The advantage of design-based methods is that populations such as socially networked hidden human populations are difficult to model realistically, and the design-based inference does not rely on modeling assumptions for properties such as unbiasedness and consistency of estimators. Design-based inference methods do rely on the design being implemented according to plan, however, and exact implementation of a given design may be a very great challenge in studies of hidden human populations. This was the motivation for the development of a range of model-based methods for inference from samples in graphs, including maximum likelihood and Bayes techniques (Thompson and Frank 2000, Chow and Thompson 2003). Assuming that the initial sample is “ignorable” in the likelihood sense (Rubin 1976), or that the design is of known form so that it can be included in the likelihood and Bayes equations, these methods work for a very wide range of link-tracing sampling procedures, including most variations of the snowball and network methods. In reality, however, the initial sample may be selected in a fashion that is anything but ignorable, with selection probabilities depending on node value, node degree, and other factors. The pervasive problem of initial sample selection in link-tracing studies has been remarked upon by Spreen (1992) among others.

The approach pursued in the present paper does not assume total control over all design possibilities, but rather seeks to work with the way samples naturally tend to get selected in networked populations, whether by ethnographers or other social scientists, members of the population themselves, or automated web crawlers. Starting with those natural selection processes, we introduce iterative modifications to obtain sampling procedures that step by step approach desired selection probabilities.

Although the underlying structure of the designs in this paper depends on Markov chains, the estimators and quantities of most interest to investigators may not in fact be Markovian. For example, while the sequence of selections of sample units may depend at each step only on the most recently selected unit, the sequence by which distinct units are added to the sample depends on all units selected thus far. For this reason, the properties of a number of alternative estimators with different designs are examined using

simulation, by repeatedly selecting samples from stochastic graph realizations and from an empirical population from a study of a people at high risk for HIV/AIDS transmission.

Random walk designs are described in section 2. Uniform and targeted walk designs are introduced in sections 3 and 4 respectively. Examples are worked in section 5, including an illustrative example using as the population a realization of a stochastic graph model and an empirical example using data from a study of a population at high risk for HIV/AIDS.

2. Random Walk

The population of interest is a graph, given by a set of N nodes with labels $U = \{1, 2, \dots, N\}$ and values $\mathbf{y} = \{y_1, \dots, y_N\}$ and an $N \times N$ matrix \mathbf{A} indicating relationships or links between nodes. An element a_{ij} of \mathbf{A} is one if there is a link from node i to node j and zero otherwise. The diagonal elements a_{ii} are assumed to be zero. For node i , the row sum $a_{i\cdot}$ is the out-degree or number of nodes to which i has a link and the column sum $a_{\cdot i}$ is the in-degree or number of nodes which link to i . With an undirected graph, the matrix \mathbf{A} is symmetric and the in-degree of any node equals its out-degree.

Let W_k denote the unit or node of the graph that is selected at the k^{th} wave. If i is the node selected at the k^{th} wave, then for wave $k+1$ one of the nodes linked from i is selected at random. Thus, $\{W_0, W_1, W_2, \dots\}$ is a Markov chain with

$$P(W_{k+1} = j | W_k = i) = a_{ij} / a_{i\cdot}. \quad (1)$$

Let \mathbf{Q} denote the transition matrix of the chain with elements $q_{ij} = P(W_{k+1} = j | W_k = i)$. The chain is a random walk in that at each step, one of the neighboring states of the present state is selected at random.

If the graph consists of a single connected component, that is, if every node of the graph is reachable from every other node by some path, then the chain is irreducible and its stationary probabilities (π_1, \dots, π_N) satisfy $\pi_j = \sum \pi_i q_{ij}$ for $j = 1, \dots, N$. In fact, with the simple random walk design in a connected undirected graph the stationary probabilities can be shown (Salganik and Heckathorn 2004) to be

$$\pi_j \propto a_{\cdot j}.$$

That is, for an undirected graph consisting of only one connected component, the long term selection frequency for any node is proportional to its in-degree, which, for a nondirected graph, equals the out-degree.

Suppose one wishes to estimate a characteristic of the population graph, such as the population mean of the node values $\mu_y = \sum_{i=1}^N y_i / N$ using data from a random walk

sample. The sample mean $\bar{y} = \sum_{i \in S} y_i$ is in general not unbiased because the value y_i of a node may be related to its degree and hence to its probability of being selected. However, one can obtain an approximately unbiased estimate by weighting each sample y -value by the reciprocal of its in-degree, assuming that that information is available from the data (Salganik and Heckathorn 2004).

2.1 Random Walk with Random Jumps

In a graph with separate components or with unconnected nodes, the simple random walk just described does not have the property that every node can be eventually reached from every other node. Without this property, the limiting distribution of the random walk is sensitive to the starting distribution, since the limiting probability for a node depends on the initial probability of starting in the component that contains that node. A modification of the design which overcomes this problem allows for a jump with small probability to a node at random from the whole graph. At each step, this random walk follows a randomly selected link with probability d and, with probability $1 - d$, jumps to another node in the graph at random or with specified probability. In the Internet search literature, d is referred to as the “damping factor”, since a value of d less than one damps the effect of the out-degree of a given node (Brin and Page 1998).

The transition probabilities for the random walk with jumps are given by

$$q_{ij} = \begin{cases} (1 - d) / N + d a_{ij} / a_{i*} & \text{if } a_{i*} > 0 \\ 1 / N & \text{if } a_{i*} = 0. \end{cases} \tag{2}$$

With the small probability $1 - d$ of a random jump at any step, the Markov chain walk can potentially reach any node in the graph from any other, so that the chain is irreducible. Further, the random jumps, which include the possibility of going to node i from node i , ensure that the chain is aperiodic so that the stationary probabilities are limiting probabilities. With $d < 1$ the stationary probability of node i is not a simple function of its own in-degree, but depends also on the stationary probabilities of the nodes that link to it.

More generally, the jumps can be made with any specified probabilities $\mathbf{p} = (p_1, \dots, p_N)$ and the probability of a jump can depend on the current state, so that the transition probabilities are

$$q_{ij} = \begin{cases} (1 - d_i) p_j + d_i a_{ij} / a_{i*} & \text{if } a_{i*} > 0 \\ 1 / N & \text{if } a_{i*} = 0. \end{cases}$$

Estimates which are approximately design-unbiased for population graph characteristics can be obtained by weighting sample values inversely proportional to the limiting Markov chain selection probabilities, but with the

additional problem that these limiting probabilities are unknown and must be estimated from the sample data (see Henzinger *et al.* 2000 for an approach to this).

For the remainder of this paper, “random walk” or “ordinary random walk” will refer to the random walk with jumps unless it is specifically stated to be a random walk without the option of jumps.

3. Uniform Walk

In this section a modification of the random walk design is proposed which leads to uniform stationary probabilities $\pi = (\pi_1, \dots, \pi_N)$.

Consider first the case of the population graph consisting of only one connected component. Let \mathbf{Q} be the transition matrix for the simple random walk with transition probabilities q_{ij} given by (1). Suppose that at step k the state of the process is i . A tentative selection is made using the transition probabilities in the i^{th} row of \mathbf{Q} . Suppose that the tentative selection is node j . If the out-degree a_{j*} of node j is less than the out-degree a_{i*} of node i , then the selection for the next wave is node j , that is, $W_{k+1} = j$. If, on the other hand, the out degree of node j is greater than the out degree of node i , then a uniform random number Z is selected from the unit interval. If $Z < a_{i*} / a_{j*}$, then $W_{k+1} = j$. Otherwise, $W_{k+1} = i$.

Using the Hastings-Metropolis method (Hastings 1970), the transition matrix for the modified walk in the connected graph is constructed with elements

$$P_{ij} = q_{ij} \alpha_{ij} \quad \text{for } i \neq j$$

and

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

where

$$\alpha_{ij} = \min \left\{ \frac{a_{i*}}{a_{j*}}, 1 \right\}.$$

With a population graph containing separate components or isolated nodes, the random walk with jumps, having transition matrix \mathbf{Q} given by (2), can be modified to give

$$P_{ij} = q_{ij} \alpha_{ij} \quad \text{for } i \neq j$$

and

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

where

$$\alpha_{ij} = \min \left\{ \frac{q_{ji}}{q_{ij}}, 1 \right\}.$$

Thus, for two mutually connected nodes i and j , the acceptance probability for a transition from i to j is

$$\alpha_{ij} = \min \left\{ \frac{(1-d)/N + d/a_{j*}}{(1-d)/N + d/a_{i*}}, 1 \right\}.$$

For a transition from an isolated unit to one in a component larger than one node, the acceptance probability is $\alpha_{ij} = 1 - d$. Other acceptance probabilities have $\alpha_{ij} = 1$. Note also that for a directed graph, the acceptance probability for following an asymmetric link would be zero.

The uniform walk is implemented, when the current state is i , by selecting a candidate next state, say j , using the transition probabilities in the i^{th} row of \mathbf{Q} . A standard uniform random number Z is selected and, if $Z < \alpha_{ij}$, the next state is j , whereas otherwise the walk stays at i for one more step.

The quantity α_{ij} with the uniform walk designs depends on the known transition probabilities of the basic random walk, so does not require estimation for implementation.

4. Targeted Walk

The same approach can be used to construct a walk having any specified stationary probabilities, for example selecting nodes with high y values with higher probabilities or selecting nodes to have probabilities strictly proportional to degree, even when the graph contains separate connected components. Let $\pi_i(y)$ denote the desired stationary selection probability for the i^{th} node as a function of its y value. For example, in a study of a hidden human population at risk for HIV/AIDS, suppose it is desired to sample injection drug users ($y_i = 1$) with twice the probability of noninjectors ($y_i = 0$). The relevant transition probabilities for the value-targeted walk, using again the Hastings-Metropolis method, are

$$P_{ij} = q_{ij}\alpha_{ij} \quad \text{for } i \neq j$$

and

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

where

$$\alpha_{ij} = \min \left\{ \frac{\pi_j(y_j)q_{ji}}{\pi_i(y_i)q_{ij}}, 1 \right\}.$$

Note that the basic transition probability is known, since it depends only on out-degree of observed nodes, the chosen probability d , and the specified ratio π_j/π_i .

For a walk in which the relative selection probability depends on y value, the ratio $\pi_j(y_j)/\pi_i(y_i)$ is specified and

$$\alpha_{ij} = \min \left\{ \frac{\pi_j(y_j)q_{ji}}{\pi_i(y_i)q_{ij}}, 1 \right\}.$$

As another example of a targeted walk, the target distribution could be to have nodes selected proportional to their out-degree, that is, the number of links out. Since the degree for an isolated node is zero, one possibility, referred to as the “degree + 1” targeted walk, simply adds one to each degree, so that $\pi_i \propto a_{i*} + 1$ is the target selection probability.

A slightly different choice, referred to simply as the degree-targeted walk, adds one only to the degree of isolated nodes, so that $\pi_i \propto \max(a_{i*}, 1)$. For a degree-targeted walk of this type, the acceptance probability for a transition between two mutually connected nodes is

$$\alpha_{ij} = \min \left\{ \frac{a_{j*}(1-d)/N + 1}{a_{i*}(1-d)/N + 1}, 1 \right\}.$$

For a transition between an isolated node and one with positive degree, the probability is

$$\alpha_{ij} = \min(a_{j*}, (1-d), 1).$$

The transition probability between two nodes each having positive degree is

$$\alpha_{ij} = \min \left\{ \frac{a_{j*}}{a_{i*}}, 1 \right\}.$$

In that case

$$\alpha_{ij} = \min \left\{ \frac{a_{j*}q_{ji}}{a_{i*}q_{ij}}, 1 \right\}.$$

Since isolated nodes, without any links to other nodes, have degree zero, to give them a positive selection probability their degree can arbitrarily be assigned the value “1” in the degree-targeted walk calculation, or the value 1 can be added to the degree of every node.

5. Nonreplacement Walk Designs

The limiting distribution results of the previous sections apply exactly to walk designs with replacements, so that the selection of nodes can proceed indefinitely through the finite population. Some of the estimators used in the examples to follow, are based however on the sequence of distinct units selected through that process. The sequence of distinct units, which in effect provides a walk sample without replacement, can add new units only until the number of distinct nodes in the sample equals that of the finite population, at which point the sample mean and the population mean coincide.

A different procedure for selecting a walk sample without replacement is to directly confine the selection of the next unit at any step from the set of units not already selected, as with the “self-avoiding random walk” (Lovász 1993). If a select-reject procedure is used as with the

targeted walks, the next selection is made from the set of units not having been tentatively selected at all, whether or not the unit was accepted.

6. Estimators Based on the Values of the Accepted Nodes

With a uniform random walk with replacement the draw-by-draw sample mean of the sequence of accepted values is asymptotically unbiased for the mean of the population, because the limiting selection probabilities are all equal. The draw-by-draw sample mean is the nominal mean including repeat values, so a node's value is weighted by the number of times it is selected. With a without-replacement design this same estimator is not precisely asymptotically unbiased because the limiting probabilities are not exactly equal. The standard variance estimator based on a within-walk sample variance is not unbiased because of the dependencies within walks. Variance estimators are examined empirically in the examples.

With a targeted walk in which the limiting probability π_i of node i is proportional to c_i , an asymptotically consistent estimator, based on the limiting probabilities, is provided by the generalized ratio estimator

$$\hat{\mu} = \frac{\sum_{s_n} y_i / c_i}{\sum_{s_n} 1 / q_i}.$$

Note that the Horvitz-Thompson estimator can not be used because the proportionality constant in the inclusion probabilities is unknown, whereas in the generalized ratio estimator it cancels out. Again the limiting probabilities on which the estimator is based hold exactly for the with-replacement design. For the without-replacement variation, the estimator is examined empirically in the examples.

7. Examples

7.1 Realized Stochastic Graph

Figure 1 depicts first a small simulated population having 60 nodes. Nodes having value $y = 1$ are colored dark and nodes with value $y = 0$ are light. The entire realization is taken to be our population of interest. The model producing the realization is a stochastic block model in which the probability of a link between any two nodes depends on the values of the nodes. Links are more likely between nodes of the same type, and the dark nodes are more highly connected than the light nodes. For example, it may be of interest to estimate the proportion of positive nodes (that is, nodes with $y = 1$) in the graph. In the population graph, 24

of the 60 nodes are positive, so the true proportion is 0.4. To the right is shown the same graph but with node sizes proportional to the random walk limiting selection probabilities. Because of the higher linkage tendencies of the positive nodes, many of them have higher than average selection probabilities.

In the bottom row of Figure 1 a random walk and a uniform walk selected from the population are shown. Each starts from the same randomly selected node, labelled "1", and proceeds until five distinct nodes are selected. The arrows show the direction of following links and a jump to a new node selected at random from the graph is shown as a dotted line. Note that the random walk backtracks from the third selected node to the second one before following a new link to the fourth sample node. From the first sample node, the uniform walk passes up the higher-probability node selected by the random walk, accepting instead another of the nodes linked to it. Either of these walks can at any time take a random jump, though in the examples illustrated only the uniform walk happens to take one, in the transition from the third to the fourth sample node.

7.2 Empirical Population

Data from a study on the heterosexual transmission of HIV/AIDS in a high-risk population in Colorado Springs (Potterat *et al.* 1993, Rothenberg *et al.* 1995) are shown in Figures 2 and 3. The 595 people interviewed in the study population are represented by the nodes of the graph, and the reported sexual relationships between the respondents are shown as links between nodes. (Additional sexual links from any of the 595 to persons who were not subsequently interviewed are not shown.) The study population includes at-risk people including injecting drug users, sex workers, their sexual and drug-use partners and other close social contacts. The node variable depicted indicates sex work, with a positive value ($y = 1$) colored dark. Only sexual links are shown, though many coincide with the drug-related links. The largest sexually connected component of the graph contains 219 of the people. The next largest connected component contains 12 people, followed by a number of components of four, three and two people. The remaining nodes represent people without reported sexual contacts within the interviewed population.

The observed pattern of this population, with one connected component very much larger than the others, has been described by researchers as not atypical of studies of hidden, at-risk populations. We are using this population solely as an empirical population from which to select samples to compare sampling designs and estimators.

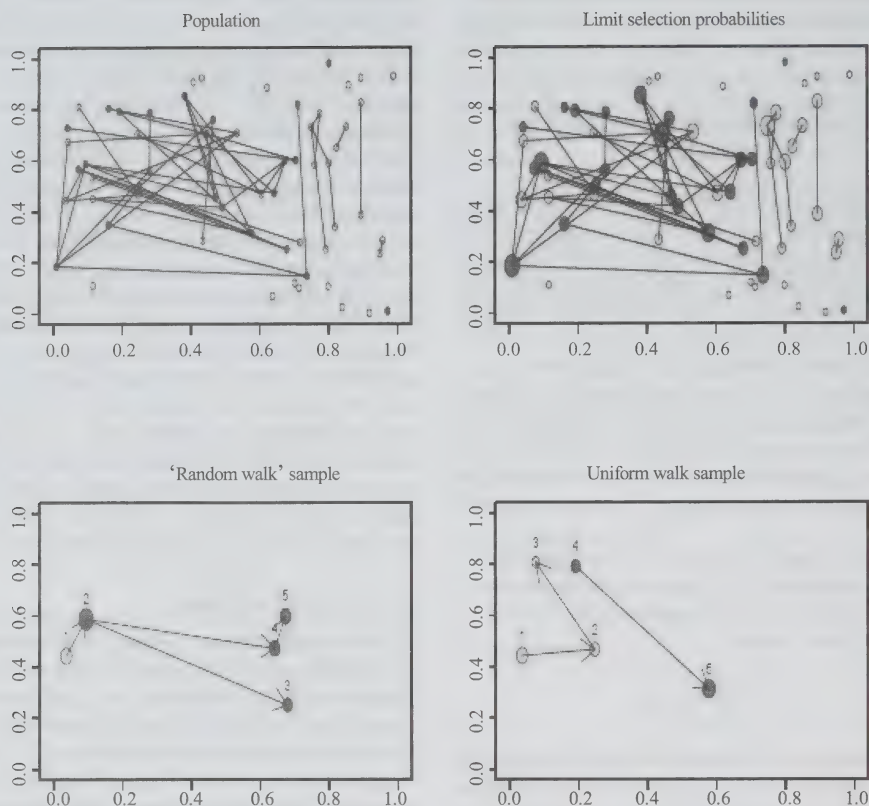


Figure 1. Top left: Population is realization of stochastic block graph model. Top right: The random walk limit probabilities of the nodes. Bottom left: Random walk of 5 steps. Bottom right: Uniform walk of 5 steps. Arbitrary axes scales are provided as a visual aid in identifying sample nodes with population nodes.

Figure 3 shows the same population with node size drawn proportional to random walk limiting selection probability.

Each plot of Figure 4 shows a cumulative sample mean of a single walk which is continued until 120 distinct nodes have been selected. The actual proportion of positive (1-valued) nodes in the empirical population (0.2235) is shown by the horizontal line in each plot.

In the top row of Figure 4, an ordinary random walk with a randomly selected starting node is shown. The left plot shows the cumulative sample mean of the distinct units. The right plot shows the same data but with the draw-by-draw sample mean, which includes repeat selections of the same node, so that each node value is weighted by the number of times that node was selected during the random walk.

In the bottom row of Figure 4 the same two types of sample mean are shown for a uniform walk that is continued until 120 distinct nodes are selected. Notice that, for the ordinary random walk, the sample mean wanders mainly above the actual mean, representing the positive bias resulting from the preferential selection of the more highly connected, high-risk people in the population. For the uniform walk, the sample mean wanders closer to the actual value, sometimes above and sometimes below. Each of these plots also gives indication of the autocorrelation present within a single Markov chain.

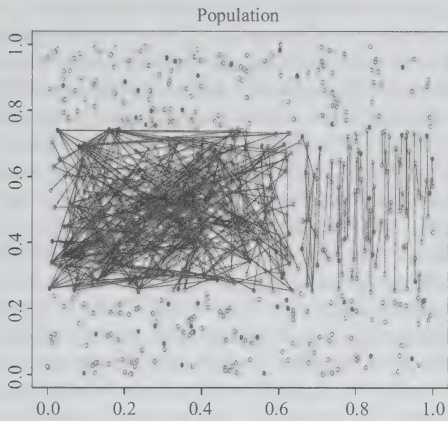


Figure 2. High-risk population in Colorado Springs study on the heterosexual transmission of HIV/AIDS (Potterat *et al.* 1993, Rothenberg *et al.* 1995, and personal communications). Dark circles represent highest-risk individuals, in this case those who have exchanged sex for money. Links shown between individuals are sexual and drug injecting partnerships.

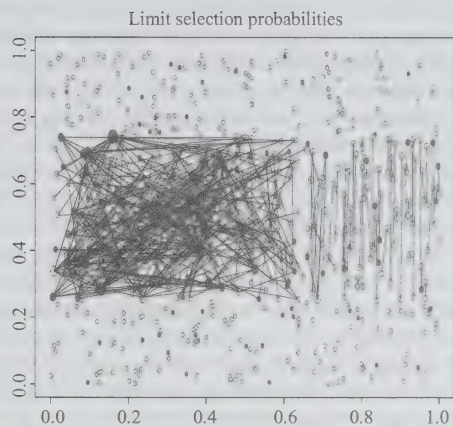


Figure 3. Limiting random walk selection probabilities for Colorado Springs population. Notice that in the real population many of the individuals with the highest-risk behavior also have high selection probabilities with the ordinary random walk, and so will tend to be overrepresented in a sample.

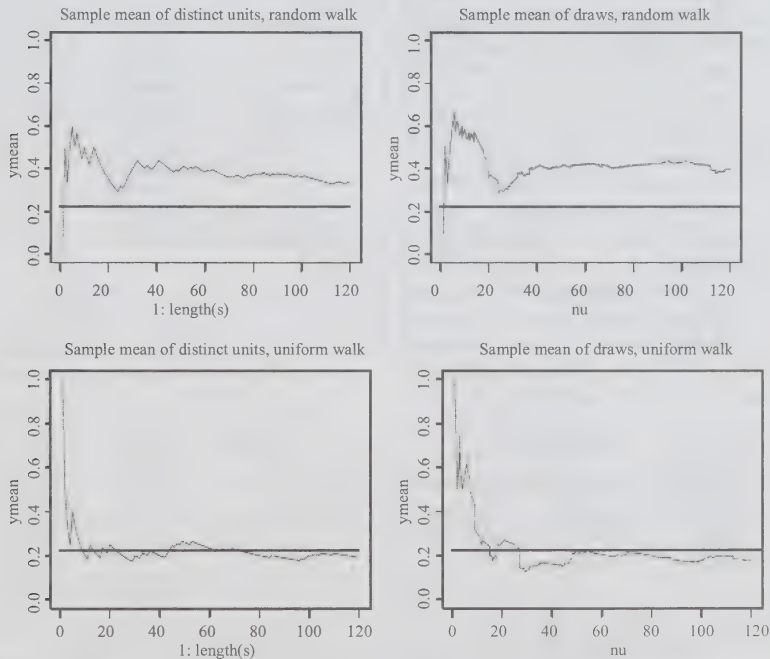


Figure 4. Sample paths of sample means for a single random walk of length 120 nodes. The top two plots are with an ordinary random walk, while the bottom two are with a uniform walk. Sample mean of the distinct units, up to the wave given by the x-axis, is plotted on the left. On the right is the sample mean of the nominal draws, so that node value is weighted by the number of times the node is selected.

The plots in Figure 5 show the expected node value as a walk progresses wave by wave, for different types of walks and with different initial distributions from which the first node is selected, for the empirical population with 595 nodes. Thus, for the k^{th} wave, the plots show $E(Y_k)$, where Y_k is the value of the node selected at the k^{th} wave. The dashed line shows the actual mean for the Colorado Springs population (0.2235). The other three lines represent three different starting distributions. In all cases, the line that starts out the lowest is the uniform initial distribution, since the mean for the initial randomly selected node equals the mean for the population. The value-dependent initial distribution, in which positive nodes ($y=1$) have twice the initial selection probability of zero nodes ($y=0$), gives the expected value line that is in all cases mostly in the middle initially and shows the strongest tendency toward initial periodicity. The degree-based initial distribution, in which initial probability of selection for a node is proportional to its degree (plus one, since isolated nodes have zero degree), forms the top line in each of the plots.

The six plots in Figure 5 show the expected values for six different types of walks. For a random walk that follows

links only, without the possibility of random jumps, the long term distribution is dependent on which component the walk starts in, which depends on the initial distribution. The three separate lines in the first figure reflect the sensitivity to the initial distribution. The random walk with jumps, on the other hand, enables any node to be reached from any other so that a limiting distribution is approached quite rapidly whatever the initial distribution. With the uniform random walk, the walk that starts with the uniform distribution stays in the uniform distribution wave after wave, and the walks that start with either of the unequal distributions depicted approach this distribution fairly rapidly. Each of the value-dependent and degree-dependent walks also approaches its limiting distribution fairly rapidly, with the expected node value considerably higher than the average node value in the population. The “degree + 1” walk approaches a distribution with selection probabilities proportional to one plus the degree for each node, while the “degree” walk has limiting probabilities proportional to the actual degree except that isolated nodes are assigned degree one.

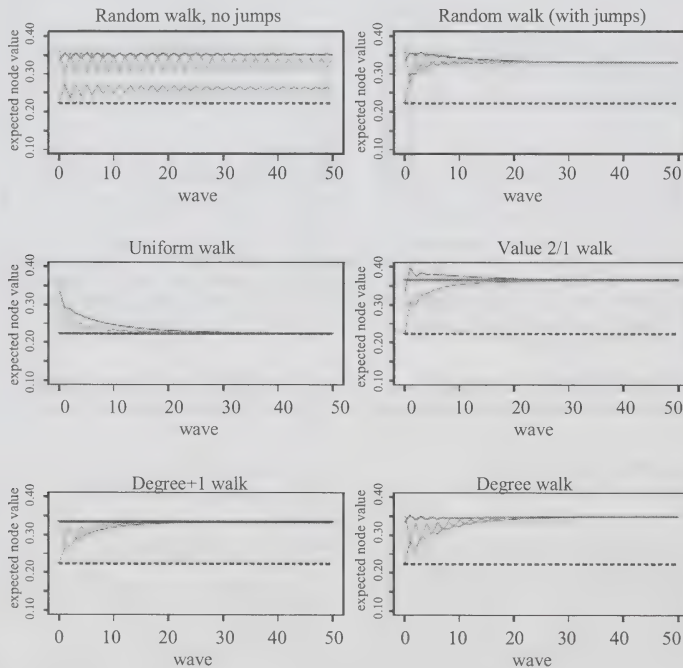


Figure 5. Expected value of node by wave for different walk designs with the Colorado Springs empirical population. Each plot shows one walk design. The dashed line is the actual mean. The other three lines show expected value for three different starting distributions. In each case the lower of the three lines starts with the uniform distribution, the middle line with the value 2/1 distribution, and the top line with the degree distribution.

Tables 1 and 2 show the calculated values of the expected value of y for the Colorado Springs study population for each type of walk, wave by wave, and with different starting distributions for the node selections. Results for ordinary random walks are in Table 1 and for uniform walks are in Table 2. The expected values are shown for the initial selections, waves 1, 2, 3, 4, 5, 6, 8, 16, and 32, and for the limit as the number of waves approaches infinity. The three initial distributions, for the selection of the first node of a walk, are random, selection in which positive nodes have twice the probability of zero-valued nodes, and selection proportional to in-degree of each node plus one. Note that, with k independent walks of a given design, the expectations at wave j would apply to the sample mean of the k y -values at wave j from each of the walks.

Table 1

Random Walks: Expected Value of y for Waves 0, 1, 2, 3, 4, 5, 6, 8, 16, 32, and Infinite. Wave 0 is the Initial Selection. Three Different Initial Selection Probability Assumptions are Used: Initial Random Selection ($\pi_0 = 1/N$ for all Nodes), Nodes with Value $y = 1$ Have Twice the Selection Probability of Nodes with Value $y = 0$ ($\pi_0 \propto y + 1$), and Initial Selection Probability Proportional to in-Degree Plus One ($\pi_0 \propto a_j + 1$). The Actual Mean of the Node Values for this Population is 0.2235294

wave	$\pi_0 = 1/N$	$\pi_0 \propto y + 1$	$\pi_0 \propto a_j + 1$
0	0.2235294	0.3653846	0.3349894
1	0.2998771	0.2752690	0.3560839
2	0.3005446	0.3587093	0.3507451
3	0.3273606	0.3082865	0.3570490
4	0.3177081	0.3594697	0.3500041
5	0.3320705	0.3179675	0.3528395
6	0.3231213	0.3542086	0.3469835
8	0.3256034	0.3490933	0.3440449
16	0.3291087	0.3372548	0.3363884
32	0.3302606	0.3313908	0.3315119
∞	0.3303787	0.3303787	0.3303787

Table 2

Uniform Walks: Expected Value of y for Waves 0, 1, 2, 3, 4, 5, 6, 8, 16, 32, and Infinite, with Three Different Initial Selection Assumptions

wave	$\pi_0 = 1/N$	$\pi_0 \propto y + 1$	$\pi_0 \propto a_j + 1$
0	0.2235294	0.3653846	0.3349894
1	0.2235294	0.2590239	0.2903147
2	0.2235294	0.2741356	0.2877974
3	0.2235294	0.2447258	0.2761270
4	0.2235294	0.2511473	0.2707929
5	0.2235294	0.2372440	0.2646280
6	0.2235294	0.2420866	0.2600923
8	0.2235294	0.2371714	0.2522952
16	0.2235294	0.2285370	0.2352150
32	0.2235294	0.2243635	0.2256228
∞	0.2235294	0.2235294	0.2235294

For the ordinary random walks, starting with the initial sample, the observed value is unbiased for the population value only for the initial selection, and thereafter the bias rapidly rises to its limiting value of 0.3303787–0.223594.

With the initial samples biased toward the positive nodes, the bias changes less as the walk progresses.

For the uniform walk, an initial random selection coincides with the stationary distribution, so that the walk continues to be unbiased wave after wave. With the initial selection in which positive nodes have twice the selection probability of zero-valued nodes, the bias is greatly reduced with each of the first few waves and the selected node values approach their unbiased limiting state. With the initial selection proportional to in-degree plus one, the bias requires a few more waves to become small. The rapid initial approach of the expected value toward the limiting value suggests that it may be desirable to have an initial “burn in” period which is not used in the estimation part. Even a very short burn in of one to three waves could substantially reduce the bias of estimators based on short walks.

Figures 6–9 show the sampling distributions of sample means and weighted estimators for different walk designs with the Colorado Springs data set. Each histogram is based on 1,000 simulations of the sampling design applied to the empirical population. For the designs in Figures 6 and 7, each sample consists of 24 walks, each having length 5, that is, continuing until 5 distinct nodes are selected. Figure 5 shows the distributions of sample means for random walks (top row) and uniform walks (bottom row). The distribution of the mean of the 24 sample means of 5 distinct units is given on the left. On the right, the mean of the 24 draw-by-draw means, incorporating repeat selections, is given.

The actual proportion (0.2235) of the y values in the empirical population is indicated by the solid triangle, while the mean of the sampling distribution is indicated by the hollow triangle. The sample means for the random walks are biased upward, while the sample means for the uniform walk are nearly unbiased. Neither is precisely unbiased, because of the way the walk continues until a fixed number of distinct nodes is selected, instead of proceeding for a fixed number of waves.

Figure 7 shows the distribution of the generalized ratio estimator for the targeted walks having stationary probabilities related to node value and to degree (node degree plus one). For comparison purposes, each of these walks was started in its own stationary distribution, in effect giving the distributions of the estimators after “burn in”. These estimators are not unbiased, since effective sample size is fixed, which affects the actual probabilities with which distinct nodes are selected in sequence, and because the denominator of the estimator is random, being the sum of the sample weights.

Figures 8 and 9 show the distributions of the same estimators and designs as in Figures 6 and 7, but with each sample consisting on one long walk of 120 distinct nodes.

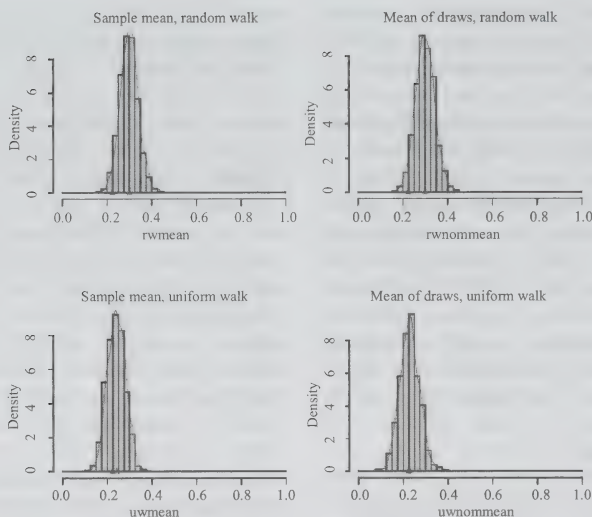


Figure 6. Distributions of sample means as estimators of the proportion of people who have exchanged sex for money in the empirical population of the Colorado Springs study, with random and uniform walks. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. Note the overestimation with sample means for ordinary random walks. Random walks are at top, uniform walks at bottom. Design was 24 walks, each of length 5, with all 120 observations used in the estimator. The number of realizations for the simulation was 1,000.

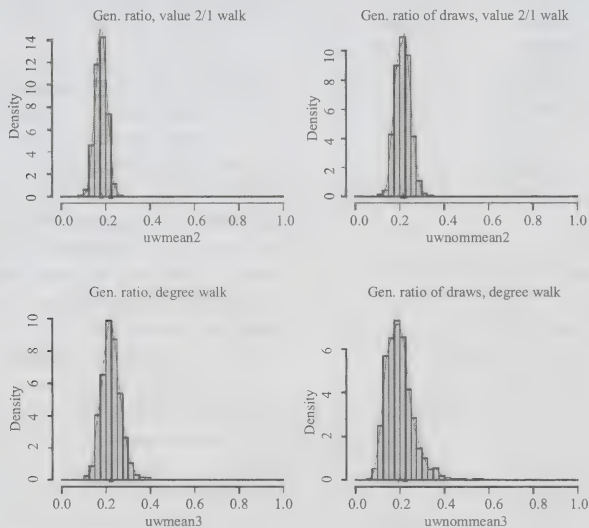


Figure 7. Distributions of generalized ratio estimators of the proportion of people who have exchanged sex for money in the empirical population of the Colorado Springs study, with targeted walks. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. Note the overestimation with sample means for ordinary random walks. Random walks are at top, uniform walks at bottom. Design was 24 walks, each of length 5, with all 120 observations used in the estimator. The number of realizations for the simulation was 1,000.

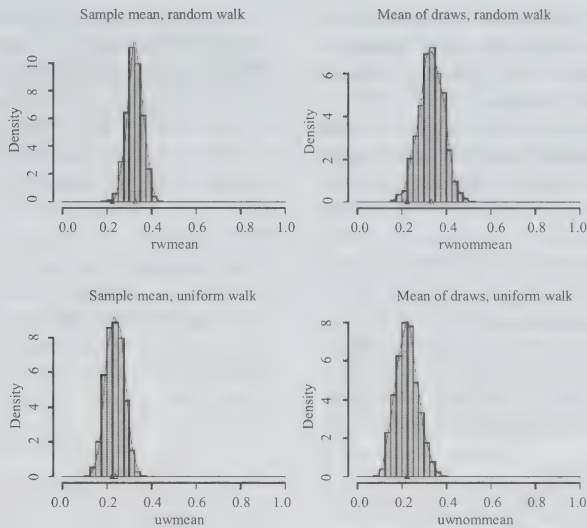


Figure 8. Distributions of sample means as estimators of the proportion of people who have exchanged sex for money in the empirical population of the Colorado Springs study, with random and uniform walks. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. Note the overestimation with sample means for ordinary random walks. Random walks are at top, uniform walks at bottom. Design was a single walk of length 120. The number of realizations for the simulation was 1,000.

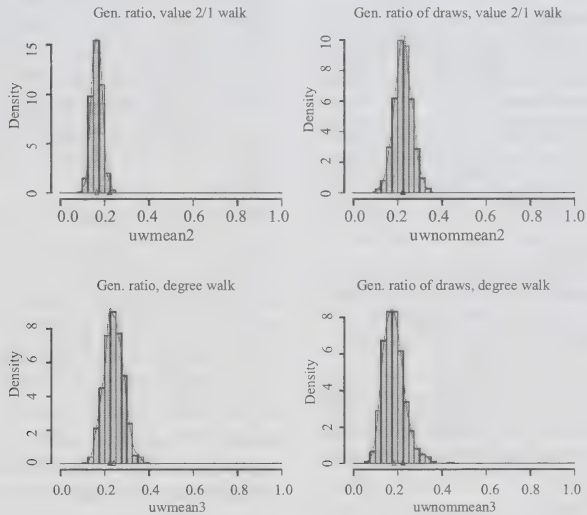


Figure 9. Distributions of generalized ratio estimators of the proportion of people who have exchanged sex for money in the empirical population of the Colorado Springs study, with targeted walks. Solid triangle is the actual proportion in the population. Hollow triangle is the mean of the distribution of the estimator. Note the overestimation with sample means for ordinary random walks. Random walks are at top, uniform walks at bottom. Design was a single walk of length 120. The number of realizations for the simulation was 1,000.

Tables 3–6 summarize the expected values and mean square errors of the estimators with the various strategies, based on the 1,000 simulation runs with the Colorado Springs data set serving as the population.

Tables 7 and 8 give the variance and the expected values of between-walk sample variances, where available, and of within-walk sample variances for the uniform walk designs.

Table 3

Means and Mean Square Errors for Sample Means of Distinct Units and Draw-by-Draw Means for Random Walks and Uniforms Walks. The Design Uses 24 Walks Each Continuing Until 5 Distinct Nodes are Included

design:	random walk	random walk	uniform walk	uniform walk
estimator:	sample mean	draw mean	sample mean	draw mean
mean	0.3008000	0.2994872	0.2423000	0.2289125
m.s.e.	0.007617465	0.007608868	0.002016378	0.001974826

Table 4

Means and Mean Square Errors for Weighted Means (Generalized Ratio Estimator), Using the Distinct Units in Each Walk or the Draw-by-Draw Selections for Value-Dependent Walks and Degree-Dependent Walks. The Design Uses 24 Walks Each Continuing Until 5 Distinct Nodes are Included

design:	value walk	value walk	degree walk	degree walk
estimator:	distinct units	draw by draw	distinct units	draw by draw
mean	0.1805114	0.2144555	0.2235257	0.1994530
m.s.e.	0.002546968	0.001195507	0.001807981	0.004382568

Table 5

Means and Mean Square Errors for Sample Means of Distinct Units and Draw-by-Draw Means for Random Walks and Uniform Walks. The Design Uses One Walk Continuing Until 120 Distinct Nodes are Included

design:	random walk	Random walk	uniform walk	uniform walk
estimator:	sample mean	draw mean	sample mean	draw mean
mean	0.3274083	0.3325171	0.2379333	0.2232534
m.s.e.	0.012004961	0.014902382	0.001777285	0.002442825

Table 6

Means and Mean Square Errors for Weighted Means (Generalized Ratio Estimator), Using the Distinct Units in Each Walk or the Draw-by-Draw Selections for Value-Dependent Walks and Degree-Dependent Walks. The Design Uses One Walk Continuing Until 120 Distinct Nodes are Included

design:	value walk	value walk	degree walk	degree walk
estimator:	distinct units	draw by draw	distinct units	draw by draw
mean	0.1652275	0.2254267	0.2404622	0.1835336
m.s.e.	0.003952703	0.001578039	0.002115518	0.003951540

Table 7

Variance of Estimators and Expected Values of Between-Walk and Within-Walk Sample Variances for the Uniform Random Walk, for the Design with 24 Walks of 5 Distinct Nodes Each

	estimator: sample mean	draw-by-draw mean
variance of estimator:	0.001665709	0.001947796
E (between-walk variance)	0.001584203	0.001919005
E (average within-walk variances)	0.001515521	0.001231983

Table 8

Variance of Estimators and Expected Values of Within-Walk Sample Variance for the Uniform Random Walk, for the Design with a Single Walk of 120 Distinct Nodes. (No Between-Walk Sample Variance is Available for this Design)

	estimator: sample mean	draw-by-draw mean
variance of estimator:	0.001571384	0.002445194
E (average within-walk variances)	0.001510515	0.001429126

Table 9

Acceptance Rates for the Uniform and Targeted Walks in the Empirical Population

	design: uniform walk	value walk	degree + 1 walk	degree walk
acceptance rate	0.62	0.60	0.85	0.88

8. Acceptance Rates

The principal advantages of the controlled Markov chain sampling designs, such as the uniform and targeted walks, are (1) they make the limiting selection probabilities known from the data so that they can be used in estimation; (2) the limiting probabilities are chosen, so that certain types of nodes or graph characteristics may be preferentially selected; (3) the estimates are design based and so certain of their key properties do not depend on assumptions, which might turn out to be incorrect, about the population graph itself; and (4) with increasing chain length, the expected values of estimates tend to move toward the corresponding graph quantities even when the initial selection distribution is different from the limiting one. Further, the uniform walk design produces a sample that, without weighting or analysis, is at face value “representative” in some respects of the larger population.

An important practical concern with the uniform and targeted walks is the acceptance rate, that is, the average probability a tentatively selected node is accepted. Tentatively selected nodes that are rejected do not contribute to the simple estimators. For a population such as the Internet, in which tentative selections and accept/reject decisions can be automated and made quickly, the acceptance rate may not be critical. Sampling simply continues until a suitable number of nodes are accepted. For studies of hidden human populations, sample sizes tend to be small. Members of the population are difficult to find and interviews may be time consuming. In some studies, however, the decision to accept or reject, based on a tentatively selected person’s out degree, may be fairly quickly ascertained through a short screening interview. Even so, it is desirable to have a sampling method with as high an acceptance rate as possible.

The random walks have acceptance probability equal to one, but do not in general have known or controlled limiting probabilities. If one thinks of the underlying random walk as the natural, uncontrolled walk through a population, then a controlled walk having a limiting distribution close to the natural random walk of the population would be expected to have a higher acceptance rate than a controlled having a limiting distribution very different from the natural random walk. That is, a controlled walk with a stationary distribution not far from the underlying random walk distribution should require less modification through the rejection of tentatively selected nodes than one with stationary distribution far from the natural random walk tendencies.

As mentioned earlier, the stationary probabilities for an ordinary random walk in a nondirected graph with a single component are proportional to the degrees of the nodes. When there is more than one connected component, the random jump innovation is necessary to ensure that every node is reachable and to produce a single stationary distribution not dependent on the starting distribution, and the limiting probabilities are influenced by, but not strictly proportional to, the node degrees. Even with the random jump innovation and the induced acceptance probabilities, the targeted walks producing stationary probabilities proportional to node degrees may be the closer than the other controlled walks under consideration to the natural random walk distribution. Indeed, in Figure 5 it is evident that, for the empirical population, the equilibrium distribution of the expected node value for the degree + 1 walk is closer to the equilibrium for the random walk with jumps than is any of the other controlled designs studied.

For the empirical population from the HIV/AIDS heterosexual transmission study, the acceptance rates for the different designs are given in Table 9. For the uniform walk design, the acceptance rate was 62 percent. For the value walk, giving twice the limiting probability for the high risk as for the low risk people, the acceptance rate was 60 percent. For the degree walk, in which the limiting probability was proportional to the degree plus one, the acceptance rate was 85 percent. For the degree walk with one added only for the degree of the isolated nodes, the acceptance rate was 88 percent.

9. Discussion

The uniform and target walk sampling designs serve to make the limiting selection probabilities known from the data so that they can be used in estimation. Further, the limiting probabilities are chosen, so that certain types of nodes or graph characteristics may be preferentially selected. Dependence on the initial selection, which may be uncontrolled, decreased step by step.

The estimators used in this paper with the uniform and targeted walk designs can be said to be design based. Even though the exact design based selection probabilities may be unknown if they are unknown in the initial selection, the stationary selection probabilities are used in the estimators. With increasing chain length, these probabilities become more accurate and the expected values of estimates move toward the corresponding graph quantities. The design based estimation methods have the advantage that certain of their properties, such as design unbiasedness or consistency, do not depend on model based assumptions that would possibly be incorrect. The design based estimates have the additional attractive quality that they are very simple and easy to understand and explain, and can even produce data that can be presented without analysis or interpretation as representative in important characteristics of the wider population of interest.

The use of Markov Chain Monte Carlo algorithms for data analysis with complicated models is common in statistics. The methods described here are unusual in that the Markov Chain methods are applied to real-world populations to actually obtain the data, with the result that the data thus obtained can be easily analyzed by hand. In fact, one could go a step farther and construct a complex Bayes stochastic graph model for the population, using Markov Chain Monte Carlo methods in the conventional fashion in analyzing the data as well as in their collection.

The uniform or targeted walk designs are useful to obtain samples of accepted nodes that have certain desirable properties in relation to the population, that provide very simple estimators of population quantities, or that could provide an initial sample for another design. It should be noted that nodes that were observed but then "rejected" under the design are actually still part of the data. Their values can still be incorporated into estimates if desired using the Rao-Blackwell method applied once the chain has reached approximate equilibrium, though the estimates then are computationally complex.

Another alternative is to use model based methods such as Bayes estimates. The model based methods require, in addition to adequate stochastic graph modeling of the population, an ignorable initial selection procedure, which is not in general satisfied with initial selections biased by node or degree values, or else adequate modeling of the non-ignorable selection procedure as part of the likelihood. Targeted walk designs producing an asymptotic distribution unrelated to the nonignorable selection procedure and hence approximately unrelated to node or degree values outside of the sample could provide the initial selections for a sample with which model based inference methods could then be applied.

Acknowledgements

Support for this work was provided by funding from the National Center for Health Statistics, the National Science Foundation (DMS-9626102 and DMS-0406229), and the National Institutes of Health (R01-DA09872). I would like to thank John Potterat and Steve Muth for advice and use of the data from the Colorado Springs study.

References

- Birnbaum, Z.W., and Sirken, M.G. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. *Vital and Health Statistics*, Serie 2, No.11. Washington: Government Printing Office.
- Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International World Wide Web Conference*, Elsevier, 107-117.
- Chow, M., and Thompson, S.K. (2003). Estimation with link-tracing sampling designs-a Bayesian approach. *Survey Methodology*, 29, 197-205.
- Felix-Medina, M.H., and Thompson, S.K. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics*, 20, 19-38.
- Frank, O. (1977). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.
- Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.
- Frank, O., and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.
- Hastings, W.K. (1970). Monte-Carlo sampling methods using Markov chains and their application. *Biometrika*, 57, 97-109.
- Heckathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.
- Heckathorn, D.D. (2002). Respondent driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.
- Henzinger, M.R., Heydon, A., Mitzenmacher, M. and Najork, M. (2000). On near-uniform URL sampling. *Proceedings of the Ninth International World Wide Web Conference*, Elsevier, 295-308.
- Klov Dahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. In *The Small World*, (Ed. M. Kochen) Norwood, NJ: Ablex Publishing, 176-210.
- Lovász, L. (1993). Random walks on graphs: A survey. In *Combinatorics, Paul Erdős is Eighty*, (Eds. D. Miklós, D. Sós and T. Szőni), János Bolyai Mathematical Society, Keszthely, Hungary, 2, 1-46.
- Potterat, J.J., Woodhouse, D.E., Rothenberg, R.B., Muth, S.Q., Darrow, W.W., Muth, J.B. and Reynolds, J.U. (1993). AIDS in Colorado Springs: Is there an epidemic? *AIDS*, 7, 1517-1521.
- Rothenberg, R.B., Woodhouse, D.E., Potterat, J.J., Muth, S.Q., Darrow, W.W. and Klov Dahl, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. In *Social Networks*, (Eds. R.H. Needle, S.G. Genser and R.T. Trotter) Drug Abuse, and HIV Transmission, NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 3-19.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Salganik, M.J., and Heckathorn, D.D. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34, 193-239.
- Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: what and why? *Bulletin de Methodologie Sociologique*, 36, 34-58.
- Thompson, S.K., and Collins, L.M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence*, 68, S57-S67.
- Thompson, S.K., and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26, 87-98.

Using Missing Data Methods to Correct for Measurement Error in a Distribution Function

Gabriele B. Durrant and Chris Skinner¹

Abstract

This paper considers the use of imputation and weighting to correct for measurement error in the estimation of a distribution function. The paper is motivated by the problem of estimating the distribution of hourly pay in the United Kingdom, using data from the Labour Force Survey. Errors in measurement lead to bias and the aim is to use auxiliary data, measured accurately for a subsample, to correct for this bias. Alternative point estimators are considered, based upon a variety of imputation and weighting approaches, including fractional imputation, nearest neighbour imputation, predictive mean matching and propensity score weighting. Properties of these point estimators are then compared both theoretically and by simulation. A fractional predictive mean matching imputation approach is advocated. It performs similarly to propensity score weighting, but displays slight advantages of robustness and efficiency.

Key Words: Donor imputation; Fractional imputation; Hot deck imputation; Multiple imputation; Nearest neighbour imputation; Predictive mean matching; Propensity score weighting.

1. Introduction

Measurement error may lead to biased estimation of distribution functions (Fuller 1995). In this paper we consider approaches to correcting for this bias when, in addition to sample observations on the erroneously measured variable, values of the accurately measured variable are available for a subsample. When the subsample is selected using a randomised scheme, the set-up is an instance of the well-studied problem of double sampling (*e.g.*, Tenenbein 1970). In this case, unbiased estimates can be constructed from the subsample alone, but use of data on the correlated surrogate variable for the whole sample may improve efficiency. See, for example, Luo, Stokes and Sager (1998). In this paper we shall suppose that the subsample is not selected by a known randomised scheme, but rather by an unknown missing data mechanism. We shall just assume that the accurate variable is missing at random (MAR) (Little and Rubin 2002), conditional on variables measured on the whole sample. Some inference methods are available for this problem if we are willing to make strong parametric assumptions about the true distribution (*e.g.*, Buonaccorsi 1990) or about the measurement error model (*e.g.*, Luo *et al.* 1998). We shall not consider such methods further, however, since we suppose that we are dealing with an application where such assumptions are unrealistic. Instead, the novel feature of this paper is to view inference in this measurement error set-up as a missing data problem and to consider the application of imputation and weighting methods from the missing data literature. Our focus will be on the choice of such methods to improve point estimation of the distribution function, in terms of bias, efficiency and robustness to model

assumptions. We shall only consider variance estimation briefly.

This paper is motivated by an application to the estimation of the distribution of hourly pay in the United Kingdom (UK), using data from the UK Labour Force Survey (LFS). In the LFS there are two ways of measuring hourly pay. The traditional method is to obtain information about earnings and hours worked and to derive a measure of hourly pay from this information. We refer to the variable derived in this way as the *derived hourly pay* variable. A more recent method of measuring hourly pay is to ask respondents directly about their hourly pay. We refer to the resulting measure of hourly pay as the *direct variable*. Skinner, Stuttard, Beissel-Durrant and Jenkins (2002) describe and provide empirical evidence of many sources of measurement error in the derived variable and conclude from their study that the direct variable measures hourly pay much more accurately than the derived variable. The problem with the direct variable is that it is missing for about 43% of all cases. The application is outlined in Section 8 and described in greater detail in Skinner *et al.* (2002), who also proposed the use of imputation to address the measurement error problem. This paper extends that work by considering a wider class of approaches to missing data and by comparing their properties both theoretically and via simulation. The imputation approach developed in this paper, which extends that considered by Skinner *et al.* (2002), has now been implemented by the UK Office for National Statistics as a new approach to producing low pay estimates.

The paper is structured as follows. The estimation problem is discussed in section 2. Imputation and weighting

1. Gabriele B. Durrant and Chris Skinner, University of Southampton, United Kingdom. E-mail: cjs@soton.ac.uk.

approaches are set out in sections 3 and 4 respectively and their properties are studied and compared theoretically in section 5 and via a simulation study in section 7. Variance estimation is considered briefly in section 6. Section 8 discusses the application of the methods to the LFS. Some concluding remarks are given in section 9.

2. The Estimation Problem

Let y_i be the (true) value of a variable of interest associated with unit i in a finite population U . The distribution function of the variable in U is:

$$F(y) = N^{-1} \sum_{i \in U} I(y_i \leq y), \quad (1)$$

where $I(\cdot)$ is the truth function ($I(E) = 1$ if E is true and $= 0$ otherwise) and y may take any specified value. Suppose that a survey is conducted on a sample $s \subset U$ and that the variable is measured as y_i^* for units $i \in s$. The difference between y_i^* and y_i represents measurement error. Suppose that the true value y_i is recorded for a subset of sample units and that we write $r_i = 1$ if y_i is recorded and $r_i = 0$ otherwise. Let x_i be a vector of auxiliary variables also recorded in the survey. Our data consist of values y_i^* , x_i and r_i for $i \in s$ and values y_i for $i \in s$ when $r_i = 1$. The problem is how to use these data to make inference about $F(y)$.

In the LFS application, the units are employees, s is the set of unit respondents in the LFS sample, y_i^* is the value of the derived hourly pay variable and y_i is the value of the direct variable for employee i . The value y_i is assumed equal to the true hourly pay.

The primary feature of this inference problem that concerns us is the missingness of y_i values and we consider two approaches to handle this missingness:

- imputation of y_i for units $i \in s$ where $r_i = 0$, using the values y_i^* and x_i as auxiliary information;
- weighting of an estimator based upon the responding subsample $s_1 = \{i \in s; r_i = 1\}$, in particular, the use of propensity score weighting (Little 1986).

These approaches to estimating $F(y)$ will be discussed in the following two sections.

Inference will be discussed under a model-based framework, in which it is assumed that the population values $(y_i, y_i^*, x_i, r_i), i \in U$, are independently and identically (IID) distributed and that sampling is ignorable, that is the distribution of (y_i, y_i^*, x_i, r_i) is the same whether or not $i \in s$. In section 8 we shall comment on how the methods developed under these assumptions may be adapted to handle the sampling design of the LFS and the use of weights to compensate for unit non-response in the survey.

3. Imputation Approaches

Suppose initially that it is possible to observe y_i for all $i \in s$. Then, under the assumptions given in the previous section,

$$\hat{F}(y) = n^{-1} \sum_{i=1}^n I(y_i < y) \quad (2)$$

would be an unbiased estimator of $F(y)$, in the sense that $E[\hat{F}(y) - F(y)] = 0$ for all y , where we write $s = \{1, \dots, n\}$ and the expectation is with respect to the model, conditional on the selected sample s . To address the problem that y_i is missing when $r_i = 0$, suppose that y_i is replaced in (2) by an imputed value y_i^I when $r_i = 0$ (and $i \in s$) and let $\tilde{y}_i = y_i$ if $r_i = 1$ and $\tilde{y}_i = y_i^I$ otherwise. The resulting estimator of $F(y)$ is

$$\tilde{F}(y) = n^{-1} \sum_{i=1}^n I(\tilde{y}_i < y). \quad (3)$$

A sufficient condition for $\tilde{F}(y)$ to be an unbiased estimator of $F(y)$ is that the conditional distribution of y_i^I given $r_i = 0$, denoted $f(y_i^I | r_i = 0)$, is the same as the conditional distribution $f(y_i | r_i = 0)$. However, since y_i is only observed when $r_i = 1$, the data provide no direct information about $f(y_i | r_i = 0)$ without further assumptions. We consider two possible assumptions.

Assumption (MAR): r_i and y_i are conditionally independent given y_i^* and x_i .

Assumption (Common Measurement Error Model): r_i and y_i^* are conditionally independent given y_i and x_i .

The first assumption is the standard one made when using imputation or weighting (Little and Rubin 2002) and is the one which we shall make. The second assumption is that the measurement error model, defined as the conditional distribution of y_i^* given y_i and x_i , is the same for respondents ($r_i = 1$) and nonrespondents ($r_i = 0$). We shall use the second assumption in the simulation study in section 7 to assess robustness of MAR-based procedures. Inference under the second assumption is more difficult, however, and appears to require stronger modelling assumptions about the distribution of y_i and x_i ; we are considering this problem in other research and do not pursue this further in this paper. The plausibility of these two assumptions for the LFS application is discussed further in Skinner *et al.* (2002).

Under the MAR assumption we have $f(y_i | y_i^*, x_i, r_i = 0) = f(y_i | y_i^*, x_i, r_i = 1)$ and a sufficient condition for $\tilde{F}(Y)$ to estimate $F(Y)$ unbiasedly is that

$$f(y_i^I | y_i^*, x_i, r_i = 0) = f(y_i | y_i^*, x_i, r_i = 1). \quad (4)$$

We therefore consider an imputation approach where the conditional distribution of y given y^* and x is ‘fitted’ to the respondent ($r_i = 1$) data and then the imputed values y_i^j are ‘drawn from’ this fitted distribution at the values y_i^* and x_i observed for the nonrespondents. Suppose that the conditional distribution $f(y_i | y_i^*, x_i, r_i = 1)$ may be represented by a parametric regression model:

$$g(y_i) = h(y_i^*, x_i; \beta) + e_i, \quad E(e_i | y_i^*, x_i) = 0 \tag{5}$$

where $g(\cdot)$ and $h(\cdot)$ are given functions and β is a vector of regression parameters. A point predictor of y_i , given an estimator $\hat{\beta}$ of β based on respondent data, is

$$\hat{y}_i = g^{-1}[h(y_i^*, x_i; \hat{\beta})]. \tag{6}$$

Using \hat{y}_i for imputation may, however, lead to serious underestimation of $F(y)$ for low values of y , since such simple regression imputation is expected to reduce the variation in $F(y)$ artificially (Little and Rubin 2002, page 64). This effect might be avoided by taking $y_i^j = g^{-1}[h(y_i^*, x_i; \hat{\beta}) + \hat{e}_i]$, where \hat{e}_i is a randomly selected empirical residual (Little and Rubin 2002, page 65). Our experience is, however, that this approach fails to generate imputed values which reproduce the ‘spiky’ behaviour of hourly pay distributions in our application and may lead to bias around these spikes. We prefer therefore to restrict attention to donor imputation methods, which set $y_i^j = y_{d(i)}$ ($r_i = 0$) for some donor respondent $j = d(i)$ for which $r_j = 1$. The imputed value from a donor will always be a genuine value and will respect the spiky behaviour in our application. The basic donor imputation method we consider is predictive mean matching (Little 1988), that is nearest neighbour imputation with respect to \hat{y}_i , defined by (6), i.e.,

impute y_i by $y_{d(i)}$
satisfying $|\hat{y}_i - \hat{y}_{d(i)}| = \min_{j: r_j = 1} |\hat{y}_i - \hat{y}_j| \tag{7}$

where $r_i = 0$ and $r_{d(i)} = 1$.

Corollary 2 of Theorem 1 of Chen and Shao (2000) then provides theoretical justification for the approximate unbiasedness of the resulting estimator $\tilde{F}(y)$ for $F(y)$, if the following four conditions hold: (i) y_i is missing at random (MAR) conditional on $z_i = g^{-1}[h(y_i^*, x_i; \beta)]$, where $\beta = \text{plim}(\hat{\beta})$, (ii) the conditional expectation of y_i given z_i is monotonic and continuous in z_i , (iii) z_i and $E(y_i | z_i)$ have finite third moments and (iv) the probability of response given z is bounded above zero. These conditions seem plausible provided: the MAR assumption above holds; the distribution of y_i only depends on y_i^* and x_i via z_i ; y_i^* is a reasonably good proxy for y_i . In addition, Chen and Shao’s (2000) result needs to be adapted for the fact that the nearest neighbour is defined with respect to $\hat{\beta}$ whereas the above conditions are with respect to β . This adaptation

seems plausible since, for a sufficiently large number of respondents, close neighbours with respect to $\hat{y}_i = g^{-1}[h(y_i^*, x_i; \hat{\beta})]$ should also be close neighbours with respect to $z_i = g^{-1}[h(y_i^*, x_i; \beta)]$.

There are thus theoretical grounds that nearest neighbour imputation with respect to \hat{y}_i will lead to an approximately unbiased estimator of $F(y)$, subject to the MAR assumption and certain additional plausible conditions. It is also of interest to consider the efficiency of $\tilde{F}(y)$. The variance of $\tilde{F}(y)$ for nearest neighbour imputation may be inflated if certain donors may be used much more frequently than others. We consider a number of approaches to reducing this variance inflation effect.

First, we may restrict the number of times that respondents are used as donors by defining imputation classes by disjoint intervals of values of \hat{y}_i and drawing donors for a recipient by simple random sampling from the class within which the recipient’s value \hat{y}_i falls. The smoothing will be greatest if we draw donors without replacement. We denote this hot deck method HDIWR or HDIWOR, depending on whether sampling is with or without replacement. A second approach is to undertake donor selection sequentially and to penalize the distance function employed for determining the nearest neighbour $d(i)$ as follows

$$|\hat{y}_i - \hat{y}_{d(i)}| (1 + \mu t_{d(i)}) = \min_{j: r_j = 1} \{|\hat{y}_i - \hat{y}_j| (1 + \mu t_j)\}, \tag{8}$$

where $\mu \in \mathbb{R}^+$ is a penalty factor, t_j is the number of times the respondent j has already been used as a donor, $r_i = 0$ and $r_{d(i)} = 1$ (Kalton 1983). A third approach is to employ repeated imputed values $y_i^{j(m)}$, $m = 1, \dots, M$, for each recipient $i \in s$ such that $r_i = 0$. The resulting estimator of $F(y)$ is $M^{-1} \sum_m \tilde{F}^{(m)}(y)$, the mean of the resulting estimators $\tilde{F}^{(m)}(y)$. We refer to the third approach as fractional imputation (Kalton and Kish 1984; Fay 1996) rather than multiple imputation (Rubin 1996), since we do not require the imputation method to be ‘proper’, that is to fulfil conditions which ensure that the multiple imputation variance estimator is consistent. We do not stipulate this requirement here because our primary objective is point estimation. In our use of fractional imputation we aim to select donors $d(i, m)$, $m = 1, \dots, M$, each a close neighbour to i , so that $\tilde{F}^{(m)}(y)$ remains approximately unbiased for $F(y)$. We consider the following variations of this approach.

- (i) The $M/2$ nearest neighbours above and below \hat{y}_i are taken, for $M = 2$ or 10, denoted NN2 and NN10 respectively.
- (ii) $M/2$ donors are selected by simple random sampling with replacement from the M respondents above and from the M respondents below \hat{y}_i , for $M = 2$ or 10, denoted NN2(4) and NN10(20) respectively.

- (iii) $M = 10$ donors are selected by simple random sampling with or without replacement from the imputation classes referred to in the HDIWR and HDIWOR methods described above. We refer to these as the HDIWR10 and HDIWOR10 methods.

For comparison we also consider the Approximate Bayesian Bootstrap method of multiple imputation (Rubin and Schenker 1986), denoted ABB10, defined with respect to the imputation classes referred to in the HDIWR and HDIWOR methods.

4. Weighted Estimation

The estimator $\tilde{F}(y)$ implied by the different imputation approaches considered in the previous section may be expressed in weighted form as:

$$\tilde{F}(y) = \sum_{i \in s_1} w_i I(y_i < y) / \sum_{i \in s_1} w_i, \quad (9)$$

where $s_1 = \{i \in s; r_i = 1\}$ is the set of respondents and $w_i = 1 + d_i / M$, where d_i is the total number of times that respondent i is used as a donor over the M repeated imputations. Note that $\sum_{i \in s_1} w_i = n$. Another choice of weight would be to set w_i equal to the reciprocal of an estimated value of the propensity score, $\Pr(r_i = 1 | y_i^*, x_i)$ (Little 1986). This approach has been proposed for the hourly pay application by Dickens and Manning (2004). The propensity score might be estimated, for example, under a logistic regression model relating r_i to y_i^* and x_i . Under the MAR assumption, the resulting estimator $\tilde{F}(y)$ will be approximately unbiased assuming validity of the model for the conditional distribution $f(r_i | y_i^*, x_i)$ and some regularity conditions, such as those described in section 3 for the imputed estimator. Note that the need to model $f(r_i | y_i^*, x_i)$ replaces the need to model $f(y_i | y_i^*, x_i)$ in the imputation approach.

5. Properties of Imputation and Weighting Approaches

In this section we investigate and compare the theoretical properties of the imputation and propensity score weighting approaches introduced in the previous two sections under various simplifying assumptions. We fix y and set $u_i = I(y_i < y)$. Letting $N \rightarrow \infty$ we suppose that the parameter of interest is $\theta = E(u_i)$. We consider the imputation approach first and suppose that y_i depends upon y_i^* and x_i only via $z_i = g^{-1}[h(y_i^*, x_i; \beta)]$ and that y_i is missing at random given z_i . Ignoring the difference between β and $\hat{\beta}$, assuming s_1 is large, we consider nearest neighbour

imputation with respect to z_i . As in (9) the imputed estimator of θ may be expressed as

$$\hat{\theta}_{\text{IMP}} = \sum_{i \in s_1} w_i u_i / \sum_{i \in s_1} w_i \quad (10)$$

where $w_i = 1 + d_i / M$ (and $\sum_{i \in s_1} w_i = n$). We write the corresponding expression for propensity score weighting as $\hat{\theta}_{\text{PS}}$ with w_i replaced by $w_{\text{PS}i}$. Let $z_{\text{PS}i}$ be the scalar function of y_i^* , x_i upon which r_i depends and write:

$$\Pr(r_i = 1 | y_i^*, x_i) = \pi(z_{\text{PS}i}). \quad (11)$$

Just as we ignored the difference between β and $\hat{\beta}$, we initially ignore error in estimating $\pi(z_{\text{PS}i})$ and write $w_{\text{PS}i} = \pi(z_{\text{PS}i})^{-1}$.

The imputation and propensity score weighting approaches may be expected to yield similar estimators if z_i and $z_{\text{PS}i}$ are similar, that is they are close to deterministic functions of each other, and M is large. To see this, consider a simple example of the imputation approach, where the donor is drawn randomly from an imputation class c of close neighbours with respect to z_i , containing m_c respondents and $n_c - m_c$ nonrespondents, as described in section 3. In this case, w_i will approach $1 + (n_c - m_c) / m_c = n_c / m_c$ as $M \rightarrow \infty$ and this is the inverse of the response rate within the class (David, Little, Samuël and Triest 1983). More generally, with the fractional nearest neighbour imputation approach considered in section 3, the weight $w_i = 1 + d_i / M$ may be interpreted as a local (with respect to z_i) nonparametric estimate of $\Pr(r_i = 1 | z_i)^{-1}$ despite the fact that imputation is based upon a model for y_i given z_i rather than r_i given z_i . Thus, the imputation approach may be expected to lead to similar estimation results to propensity score weighting if z_i and $z_{\text{PS}i}$ are deterministic functions of each other. In general, however, this will not be the case. Since $\Pr(r_i = 1 | z_i)$ may be expressed as an average of $\Pr(r_i = 1 | y^*, x)$ across values of y^* and x for which $z = z_i$, we may interpret w_i as a smoothed version of $w_{\text{PS}i}$ and may expect it to show less dispersion. This suggests that it may be possible to use imputation to improve upon the efficiency of estimates based on propensity score weighting, as also discussed by David *et al.* (1983) and Rubin (1996, section 4.6). To investigate this further, assuming MAR and the other assumptions in sections 3 and 4 upon which the approaches are based, both imputation and weighting approaches lead to approximately unbiased estimation of $F(y)$ and we may focus our comparison on relative efficiency.

It follows from equation (3.3) of Chen and Shao (2000) that the variance of $\hat{\theta}_{\text{IMP}}$ may be approximated for large n by

$$\text{var}(\hat{\theta}_{\text{IMP}}) \approx n^{-2} E \left[\sum_{i \in s_1} w_i^2 V(u_i | z_i) \right] + n^{-1} V[\psi(z_i)], \quad (12)$$

where $\psi(z_i) = E(u_i | z_i)$ and any impact of estimating β is ignored. Note that Chen and Shao (2000) consider single imputation with $M = 1$ but their proof of this result carries through if $M > 1$. It is convenient to reexpress this result as

$$\text{var}(\hat{\theta}_{\text{IMP}}) \approx n^{-1}\sigma^2 + n^{-2}E\left[\sum_{s_i}(w_i^2 - w_i)V(u_i | z_i)\right], \quad (13)$$

using the identity

$$V[\psi(z_i)] = \sigma^2 - E[V(u_i | z_i)], \quad (14)$$

where $\sigma^2 = V(u_i)$ and a corollary of Chen and Shao's (2000) Theorem 1 that

$$E\left[n^{-1}\sum_{s_i}w_iV(u_i | z_i)\right] = E[V(u_i | z_i)] + o_p(n^{-1/2}). \quad (15)$$

Note that $w_i^2 - w_i = (d_i/M)(1 + d_i/M) \geq 0$. Expression (13) may be interpreted from both 'missing data' and 'measurement error' perspectives. From a missing data perspective, the first term in (13) is just the variance of $\hat{\theta}$ in the absence of missing data and the second term represents the inflation of this variance due to imputation error. From a measurement error perspective, we may consider limiting properties under 'small measurement error asymptotics' (Chesher 1991), that is where $y_i^* \rightarrow y_i$ and $V(u_i | z_i)$ approaches zero. In this case, the second term also approaches zero and $\hat{\theta}_{\text{IMP}}$ becomes 'fully efficient', i.e., its variance approaches σ^2/n .

Let us now consider propensity score weighting. We make the corresponding assumption that y_i is missing at random given z_{PSi} . Linearising the ratio in (9), with w_{PSi} in place of w_i , using the fact that $E(\sum_{s_i}w_{\text{PSi}}) = n$ and initially ignoring the impact of estimating the propensity score we may write

$$\begin{aligned} \text{var}(\hat{\theta}_{\text{PS}}) &\approx n^{-2} \text{var}\left[\sum_{s_i}w_{\text{PSi}}(u_i - \theta)\right] \\ &= n^{-1} E[w_{\text{PSi}}(u_i - \theta)^2], \end{aligned} \quad (16)$$

which may be expressed alternatively as

$$\begin{aligned} \text{var}(\hat{\theta}_{\text{PS}}) &\approx n^{-2}E\left[\sum_{s_i}w_{\text{PSi}}^2V(u_i | z_{\text{PSi}})\right] \\ &\quad + n^{-1}E\{w_{\text{PSi}}[\psi(z_{\text{PSi}}) - \theta]^2\} \end{aligned} \quad (17)$$

To compare the efficiency of weighting and imputation it is convenient to use (14) and (15) (which hold also with w_{PSi} in place of w_i) to obtain

$$\begin{aligned} \text{var}(\hat{\theta}_{\text{PS}}) &\approx n^{-1}\sigma^2 \\ &\quad + n^{-2}E\left[\sum_{s_i}(w_{\text{PSi}}^2 - w_{\text{PSi}})V(u_i | z_{\text{PSi}})\right] \\ &\quad + n^{-1}E\left\{\sum_{s_i}[w_{\text{PSi}} - 1][\psi(z_{\text{PSi}}) - \theta]^2\right\}. \end{aligned} \quad (18)$$

Note that, in comparison with (13), this involves a third term, which does not necessarily converge to zero as $y_i^* \rightarrow y_i$ and $V(u_i | z_{\text{PSi}}) \rightarrow 0$. Hence propensity score weighting does not become fully efficient as the measurement error disappears. The second term of (18) may also be expected to dominate the second term of (13) when $V(u_i | z_i)$ and $V(u_i | z_{\text{PSi}})$ are constant and equal, since, recalling that $\sum_{s_i}w_i = E(\sum_{s_i}w_{\text{PSi}}) = n$, these second terms are primarily determined by the variances of the weights w_i and w_{PSi} , and, provided M is sufficiently large, we may expect w_i to display less variation than w_{PSi} , as argued above.

The above discussion ignores the potential impact of estimating β or estimating a parameter vector α upon which the propensity score $\Pr(r_i = 1 | y_i^*, x_i)$ may be assumed to depend. Kim (2004) shows in fact that the estimation of α by its maximum likelihood estimator $\hat{\alpha}$ reduces the variance of $\hat{\theta}_{\text{PS}}$ as follows:

$$\begin{aligned} \text{var}(\hat{\theta}_{\text{PS}}) &\approx \text{var}(\tilde{\theta}_{\text{PS}}) \\ &\quad - \text{cov}(\tilde{\theta}_{\text{PS}}, \hat{\alpha}) \text{var}(\hat{\alpha})^{-1} \text{cov}(\hat{\alpha}, \tilde{\theta}_{\text{PS}}), \end{aligned} \quad (19)$$

where $\tilde{\theta}_{\text{PS}}$ is the estimator $\hat{\theta}_{\text{PS}}$ with the estimated propensity scores replaced by their true values and where the left hand sides of (16), (17) and (18) should now be $\text{var}(\tilde{\theta}_{\text{PS}})$. We conclude from this fact and the previous discussion that, in general, $\hat{\theta}_{\text{IMP}}$ is not necessarily more efficient than $\hat{\theta}_{\text{PS}}$ or vice versa and we look to the simulation study in section 7 for numerical evidence. However, our conclusion that $\hat{\theta}_{\text{IMP}}$ is more efficient as measurement error disappears and $y_i^* \rightarrow y_i$ remains valid even in the presence of estimation error in α and β , since the impact of estimation error in β will disappear in this case with $z_i \rightarrow y_i^*$ whereas the second term in (19) when added to expression (18) will not in general reduce $\text{var}(\hat{\theta}_{\text{PS}})$ to σ^2/n in this case.

Let us finally consider the impact of departures from the MAR assumption. Under small measurement error asymptotics where $y_i^* \rightarrow y_i$ and $V(u_i | z_i) \rightarrow 0$ so $y_i^* \rightarrow y_i$, the imputation approach will provide consistent inference about θ even if the MAR assumption fails. This is not the case for the propensity score weighting approach. This suggests that the imputation approach may display more robustness to departures from the MAR assumption if the amount of measurement error is relatively small.

6. Variance Estimation

Although point estimation is the primary focus of this paper, we do now consider linearization variance estimation briefly. For propensity score weighting we refer to Kim (2004). For the single and fractional imputation methods in section 3 based upon nearest neighbour imputation, we may

consider a simplified approach based on the IID assumption set out in section 2 and the expression for the variance of $\hat{\theta}_{\text{IMP}}$ in (13).

The simple estimator of the first term σ^2/n :

$$n^{-1}\hat{\sigma}^2 = n^{-2} \sum_s w_i(u_i - \hat{\theta}_{\text{IMP}})^2 \quad (20)$$

is approximately unbiased from Corollary 1 of Chen and Shao (2000). It follows that an approximately unbiased estimator of $\text{var}(\hat{\theta}_{\text{IMP}})$ is

$$\hat{V}(\hat{\theta}_{\text{IMP}}) = n^{-1}\hat{\sigma}^2 + n^{-2} \sum_s (w_i^2 - w_i) \hat{V}(u_i | z_i) \quad (21)$$

if we can construct an approximately unbiased estimator $\hat{V}(u_i | z_i)$ of $V(u_i | z_i)$. Various approaches to estimating $V(u_i | z_i)$ seem possible. Following Fay (1999), we might consider the sample variance of u_i values for responding neighbours near to i with respect to z . An alternative approach would be to consider a model-based approach in which a model is fitted to $\psi(z_i) = E(u_i | z_i)$ for $i \in s$ giving $\hat{\psi}(z_i)$ and we set $\hat{V}(u_i | z_i) = \hat{\psi}(z_i)[1 - \hat{\psi}(z_i)]$. We have considered nonparametric methods of fitting $\psi(z_i)$, but have found with the LFS data that these lead to very similar values of $\hat{V}(\hat{\theta}_{\text{IMP}})$ as a logistic regression model for $\psi(z_i)$.

It may be possible to apply ideas in Chen and Shao (2001) or Kim and Fuller (2002) to extend the above approach to handle survey weights and a complex design. See Rancourt (1999) and Fay (1999) for other variance estimation approaches for nearest neighbour imputation and Little and Rubin (2002) for multiple imputation approaches.

7. Simulation Study

The aim of the study is to generate independent repeated samples $s^{(h)}$, $h = 1, \dots, H$, with values $y_i, y_i^*, x_i, r_i, i \in s^{(h)}$ which are realistic in relation to the LFS application, considered further in section 8, to compute the corresponding estimates $\hat{F}^{(h)}(y)$ for alternative approaches to missing data and values of y and to assess the performance of the estimators $\hat{F}(y)$ empirically. In order to employ realistic values, the samples $s^{(h)}$ of size n were drawn with replacement (*i.e.*, using the bootstrap) from an actual sample of about 16,000 employees for the March–May 2000 quarter of the LFS (only main jobs of employees aged 18+ were considered and the very small number of cases with missing values on y_i^* or x_i were omitted). The values of x_i for each sample $s^{(h)}$ were taken directly from the values in the LFS sample. Variables were chosen for inclusion in x_i if they were either related to hourly pay, measurement error in y_i^* or response r_i (see Skinner *et al.* 2002) and included for example age, gender, household position, qualifications, occupation, duration of employment, full-time/part-time,

industry and region (several of these variables were represented by dummy variables). We set $n = 15,000$, such that each $s^{(h)}$ was of a similar size as the original LFS sample, and $H = 1,000$. The values of y_i, y_i^* and r_i for each sample $s^{(h)}$ were generated from models, rather than directly from the LFS data, for the following reasons.

- y_i : these values were generated from a model because they were frequently missing in the LFS. A linear regression model was used, relating $\ln(y_i)$ to $\ln(y_i^*)$ and x_i with a normal error and with 20 covariates including squared terms in $\ln(y_i^*)$ and age and interactions between $\ln(y_i^*)$ and 5 components of x_i . The model was fitted to the roughly 7,000 cases where y_i was observed.
- y_i^* : these values were generated from a model to avoid duplicate values of (y_i^*, x_i) within each $s^{(h)}$, which it was considered might lead to an unrealistic distribution of distances between units for the nearest neighbour method. The model was a linear regression model relating $\ln(y_i^*)$ to x_i with a normal error and with 12 covariates, including a squared term in age and one interaction, fitted to the LFS data.
- r_i : these values were generated from a model to ensure that the missing data mechanism was known. Several models were fitted. The only one reported here is a logistic regression relating $\ln(y_i^*)$ to $\ln(y_i^*)$ and x_i with 17 covariates including squared $\ln(y_i^*)$ and interactions between $\ln(y_i^*)$ and two covariates. The model was fitted to the LFS data. The missing data mechanism is MAR given the y_i^* and x_i for all the results presented except those in Table 5.

Estimates $\hat{\theta}_t^{(h)}$ of two parameters ($t = 1, 2$) were obtained for each sample $s^{(h)}$,

- θ_1 = proportion with pay below the national minimum wage (= £3.00 per hour aged 18–21, £3.60 per hour aged 22+)
- θ_2 = proportion with pay between minimum wage and £5/hour.

The true values are $\theta_1 = 0.056$ and $\theta_2 = 0.185$. The bias and standard error were estimated as $\text{bias}(\hat{\theta}_t) = \bar{\theta}_t - \theta_t$ and $\text{s.e.}(\hat{\theta}_t) = [H^{-1} \sum_{h=1}^H (\hat{\theta}_t^{(h)} - \bar{\theta}_t)^2]^{1/2}$, where $\bar{\theta}_t = H^{-1} \sum_h \hat{\theta}_t^{(h)}$.

For the fractional imputation methods several different values for M were explored and $M = 10$ or 20 were chosen to achieve an increase in the efficiency whilst still being able to define a nearest neighbour imputation sensibly.

We first compare results for the alternative imputation approaches. Table 1 presents estimates of the biases of estimators of θ_1 and θ_2 for different imputation methods, for a MAR missing data mechanism. There is no evidence of significant biases for any of the nearest neighbour (NN) methods. The bias/standard error ratios are small and may be expected to be even smaller for estimates within domains *e.g.*, regions or age groups. We conclude that there is no evidence of important bias for these methods, provided the MAR mechanism holds and the model is correctly specified.

There is some evidence of statistically significant biases for each of the three methods based on imputation classes (HDIWR10, HDIWOR10, ABB10) perhaps because of the width of the classes, although the bias appears to be small relative to the standard error. Given the additional disadvantage of these methods, that the specification of the boundaries of the classes is arbitrary, these methods appear to be less attractive than the nearest neighbour methods. This finding contrasts with the preference sometimes expressed (*e.g.*, Brick and Kalton 1996, page 227) for stochastic methods of imputation, such as the HDI methods, compared to deterministic methods, such as nearest

neighbour imputation, when estimating distributional parameters.

Corresponding estimates of standard errors are given in Table 2. We find as expected that the greatest standard error occurs for the single NN1 imputation method. The variance is reduced by around 10% using the penalty function method (NN1P). About 10–20% reduction arises from using two imputations (NN2 or NN2 (4)) and around 20% reduction from using ten imputations (NN10, NN10 (20)), HDIWR10, HDIWOR10, ABB10). For a given number of imputations (2 or 10) there seem to be no obvious systematic effects of using a stochastic method (NN2 (4) or NN10 (20)) versus a deterministic method (NN2 or NN10). We would expect the standard errors for HDIWR10 to be no less than HDIWOR10, which is the case for $\hat{\theta}_1$ in table 2. The slight reduction for the standard error of estimator $\hat{\theta}_2$ is likely to be caused by a comparatively small number of simulation iterations ($H = 1,000$), which may not be fully sufficient for standard error estimation. We conclude that NN10 is the most promising approach, avoiding the bias of the imputation class methods and having appreciable efficiency gains over the methods generating one or two imputations.

Table 1
Simulation Estimates of Biases of Estimators of θ_1 and θ_2 for Different Imputation Methods,
Assuming MAR and Correct Covariates ($H = 1,000$)

Imputation Method	Bias of $\hat{\theta}_1$	Rel. Bias of $\hat{\theta}_1$	Bias of $\hat{\theta}_2$	Rel. Bias of $\hat{\theta}_2$
NN1	1.2*10 ⁻⁴ (0.9*10 ⁻⁴)	0.2 %	0.9*10 ⁻⁴ (1.7*10 ⁻⁴)	0.0 %
NN1P ¹	4.4*10 ⁻⁴ (2.6*10 ⁻⁴)	0.8 %	0.3*10 ⁻⁴ (5.1*10 ⁻⁴)	0.0 %
NN2	0.6*10 ⁻⁴ (0.8*10 ⁻⁴)	0.1 %	1.6*10 ⁻⁴ (1.5*10 ⁻⁴)	0.0 %
NN2(4)	1.4*10 ⁻⁴ (0.9*10 ⁻⁴)	0.2 %	-2.5*10 ⁻⁴ (1.5*10 ⁻⁴)	-0.1 %
NN10	0.2*10 ⁻⁴ (0.8*10 ⁻⁴)	0.0 %	-1.2*10 ⁻⁴ (1.5*10 ⁻⁴)	-0.1 %
NN10(20)	0.2*10 ⁻⁴ (0.8*10 ⁻⁴)	0.0 %	0.7*10 ⁻⁴ (1.5*10 ⁻⁴)	0.0 %
HDIWR10	2.8*10 ⁻⁴ (0.8*10 ⁻⁴)	0.5 %	26.2*10 ⁻⁴ (1.5*10 ⁻⁴)	1.4 %
HDIWOR10	2.5*10 ⁻⁴ (0.8*10 ⁻⁴)	0.4 %	28.0*10 ⁻⁴ (1.5*10 ⁻⁴)	1.5 %
ABB10	4.6*10 ⁻⁴ (0.8*10 ⁻⁴)	0.8 %	29.8*10 ⁻⁴ (1.5*10 ⁻⁴)	1.6 %

Standard errors of bias estimates are below the estimates in parentheses.

¹ Note: $H = 100$ iterations were used due to computing time.

Table 2
Simulation Estimates of Standard Errors of Estimators of θ_1 and θ_2 for Different Imputation Methods, Assuming MAR and Correct Covariates ($H = 1,000$)

Imputation Method	s.e.($\hat{\theta}_1$)	s.e.($\hat{\theta}_2$)	$\frac{V(\hat{\theta}_1)}{V_{NN1}(\hat{\theta}_1)}$	$\frac{V(\hat{\theta}_2)}{V_{NN1}(\hat{\theta}_2)}$
NN1	2.79×10^{-3}	5.43×10^{-3}	1	1
NN1P ²	2.60×10^{-3}	5.15×10^{-3}	0.87	0.91
NN2	2.68×10^{-3}	5.05×10^{-3}	0.91	0.86
NN2(4)	2.73×10^{-3}	4.88×10^{-3}	0.94	0.80
NN10	2.56×10^{-3}	4.88×10^{-3}	0.83	0.81
NN10(20)	2.57×10^{-3}	4.79×10^{-3}	0.84	0.77
HDIWR10	2.52×10^{-3}	4.66×10^{-3}	0.82	0.74
HDIWOR10	2.48×10^{-3}	4.72×10^{-3}	0.78	0.76
ABB10	2.63×10^{-3}	4.87×10^{-3}	0.88	0.80

² Note: $H = 100$ iterations were used due to computing time.

We next compare the NN10 imputation approach with propensity score weighting (PSW). We consider not only the case when the specification of the model used for imputation or weighting corresponds to the model used in the simulation, as in Table 1, but also some cases of misspecification. To ensure a fair comparison of weighting and imputation we use the same covariates when fitting both the models generating y_i and r_i . We first consider the estimated biases in Table 3. When the model for imputation (NN10) or the propensity scores is correctly specified neither method demonstrates any significant bias in the estimation of θ_1 or θ_2 . Significant bias does arise, however, in both cases if the model is misspecified by failing to include covariates used in the simulation. The amount of bias is, however, noticeably greater for the weighting approach. For example, for the estimator $\hat{\theta}_1$ the bias is 3–7 times higher under PSW than under NN10 depending on the misspecification. The impact of the misspecification seems higher for estimator $\hat{\theta}_2$, in particular for the PSW method. For this estimator, we found a 6–15 times higher bias for PSW than for NN10.

Corresponding estimated standard errors of $\hat{\theta}_1$ and $\hat{\theta}_2$ are given in Table 4. These also tend to be greater for the weighting approach, showing an increase between 5–15% in comparison to the imputation method. The increase in the standard error is higher for the second estimator $\hat{\theta}_2$, ranging from 12–15%, whereas for estimator $\hat{\theta}_1$ the increase is between 5–12%, depending on the misspecification. Consequently, the mean squared error is also higher for the weighting approach, with the increase ranging from

20% to 28% for the six values in Table 4. At least under the MAR assumption, the NN10 imputation approach appears to be preferable to propensity score weighting in terms of bias and variance.

Finally, we compare the properties of imputation (NN10) and propensity score weighting when the MAR assumption fails. We now simulate missingness according to the Common Measurement Error model assumption of section 3. The same logistic model with the same coefficients as in the previous simulation is used except that y_i^* is replaced as a covariate by y_i . Simulation estimates of biases and standard errors are presented in Table 5. We observe a non-negligible significant relative bias of around 5% for the imputation approach and a little higher for the propensity score weighting approach. The positive direction of the bias of $\hat{\theta}_1$ is as expected from arguments in Dickens and Manning (2004) and Skinner *et al.* (2002). MAR-based methods will tend to overestimate numbers of the low paid, if the CME assumption holds. This is because employees with observed y_i values tend to be lower paid than employees with missing y_i values and a MAR-based imputation method, even conditional on other variables, would tend to impute lower hourly pay values than would be the case under CME which allows for the dependency on true hourly pay. While the direction of the effect may be anticipated, the magnitude of the effect is of some importance for the robustness of MAR-based methods. The relative bias of 5% of the NN10 approach does not, however, appear to make the resulting estimates unusable.

Table 3
Simulation Estimates of Biases of Estimators of θ_1 and θ_2 for Nearest Neighbour Imputation (NN10) and Propensity Score Weighting, Assuming MAR and Correct and Misspecified Covariates ($H = 1,000$)

Method	Assumed Covariates	Bias of $\hat{\theta}_1$	Rel. Bias of $\hat{\theta}_1$	Bias of $\hat{\theta}_2$	Rel. Bias of $\hat{\theta}_2$
NN10	M1 (correct)	$-0.18*10^{-4}$	-0.03%	$-5.8*10^{-4}$	-0.31%
		$(0.64*10^{-4})$		$(1.20*10^{-4})$	
	M2	$-1.31*10^{-4}$	-0.24%	$-4.74*10^{-4}$	-0.25%
		$(0.65*10^{-4})$		$(1.23*10^{-4})$	
	M3	$-1.66*10^{-4}$	-0.30%	$-10.6*10^{-4}$	-0.57%
		$(0.63*10^{-4})$		$(1.23*10^{-4})$	
Propensity Score Weighting	M1 (correct)	$0.15*10^{-4}$	0.03%	$-2.62*10^{-4}$	-0.14%
		$(0.72*10^{-4})$		$(1.35*10^{-4})$	
	M2	$-8.96*10^{-4}$	-1.64%	$70.2*10^{-4}$	3.80%
		$(0.68*10^{-4})$		$(1.40*10^{-4})$	
	M3	$-5.02*10^{-4}$	-0.92%	$67.8*10^{-4}$	3.66%
		$(0.68*10^{-4})$		$(1.41*10^{-4})$	

Note: M1 is the correct model
M2 excludes the interactions and the square terms from the correct model
M3 drops further covariates from model M2.

Table 4
Simulation Estimates of Standard Errors of Estimators of θ_1 and θ_2 for Nearest Neighbour Imputation (NN10) and Propensity Score Weighting, Assuming MAR and Correct and Misspecified Covariates ($H = 1,000$)

Method	Assumed Covariates	s.e.($\hat{\theta}_1$)	s.e.($\hat{\theta}_2$)	MSE($\hat{\theta}_1$)	MSE($\hat{\theta}_2$)
NN10	M1 (correct)	$2.02*10^{-3}$	$3.80*10^{-3}$	$4.10*10^{-6}$	$1.49*10^{-5}$
	M2	$2.06*10^{-3}$	$3.88*10^{-3}$	$4.29*10^{-6}$	$1.54*10^{-5}$
	M3	$2.01*10^{-3}$	$3.89*10^{-3}$	$4.10*10^{-6}$	$1.63*10^{-5}$
Propensity Score Weighting	M1 (correct)	$2.27*10^{-3}$	$4.27*10^{-3}$	$5.16*10^{-6}$	$1.83*10^{-5}$
	M2	$2.17*10^{-3}$	$4.42*10^{-3}$	$5.51*10^{-6}$	$6.90*10^{-5}$
	M3	$2.16*10^{-3}$	$4.46*10^{-3}$	$4.94*10^{-6}$	$6.59*10^{-5}$

Table 5
Simulation Estimates of Biases and Standard Errors of Estimators of θ_1 and θ_2 for Nearest Neighbour Imputation (NN10) and Propensity Score Weighting. Under the (non-MAR) Common Measurement Error Model ($H = 1,000$)

Method	Bias of $\hat{\theta}_1$	Rel. Bias of $\hat{\theta}_1$	Bias of $\hat{\theta}_2$	Rel. Bias of $\hat{\theta}_2$	s.e.($\hat{\theta}_1$)	s.e.($\hat{\theta}_2$)
NN10	$29.0*10^{-4}$	5.1%	$92.0*10^{-4}$	5.0%	$2.53*10^{-3}$	$4.70*10^{-3}$
	$(0.8*10^{-4})$		$(1.48*10^{-4})$			
Propensity Score Weighting	$32.3*10^{-4}$	5.7%	$100*10^{-4}$	5.7%	$2.31*10^{-3}$	$4.42*10^{-3}$
	$(0.73*10^{-4})$		$(1.40*10^{-4})$			

8. Application to the Labour Force Survey

In this section we consider the application of the methods developed in sections 2 – 4 to LFS data. The LFS provides an important source of estimates of the distribution of

hourly pay in the UK (Stuttard and Jenkins 2001). It is a quarterly survey of households selected from a national file of postal addresses with equal probabilities by stratified systematic sampling. All adults in selected households are included in the sample. The resulting sample is clustered by

household membership but not by geography. Each selected household is retained in the sample for interview on five successive quarters and then rotated out and replaced. Questions relating to hourly pay are asked in just the first and fifth interviews, generating data on this topic for about 16,000 employees per quarter.

Two measures of hourly pay are constructed, as outlined in Section 1. The derived hourly pay variable in the LFS is defined as follows: (a) employees are asked questions about their main job to determine earnings over a reference period, (b) questions are asked to determine hours worked over the reference period and (c) the result of (a) is divided by the result of (b). The direct variable is obtained by first asking whether the respondent is paid a fixed hourly rate and then, if the answer is positive, by asking respondents what this (basic) rate is. Skinner *et al.* (2002) discuss how the derived variable suffers from many sources of measurement error, as in similar surveys in other countries (Rodgers, Brown and Duncan 1993; Moore, Stinson and Welniak 2000). They conclude that the direct variable measures hourly pay much more accurately. A working assumption in this application is that the direct variable measures hourly pay without error. The problem with the direct variable, however, is that it is missing for respondents who state that they are not paid at a fixed hourly rate (and for item nonrespondents) and this missingness is positively associated with hourly pay. The proportion of LFS respondents with a (main) job who provide a response to the direct question is about 43%. This proportion tends to be higher for lower paid employees, for example the rate is 72% among those in the bottom decile of the derived variable. The direct variable is not collected for second (and further) jobs and we therefore restrict attention only to main jobs. The aim is to use the missing data methods developed in this paper to correct for the measurement error in hourly pay. Skinner *et al.* (2002) discuss the plausibility of the two missing data assumptions in section 3 for this application.

The methods in sections 2–4 were developed under the assumption of an IID model and ignorable sampling. Employees are selected with equal probabilities in the LFS so the sampling may be viewed as ignorable with respect to the bias of point estimation but unit non-response is likely to be differential and survey weights are constructed to compensate for this non-response (ONS 1999). We propose to incorporate these survey weights into the estimator in (3) or equivalently to multiply the weights w_i in (9) by the survey weights. This is analogous to the way the pseudo-likelihood approach (Skinner 1989) weights estimators based upon an IID assumption. The aim is to use the methods of sections 2–4 to compensate for bias due to measurement error and

item non-response and the survey weights to compensate for bias due to sampling and unit nonresponse. We have not attempted to take account of the weights in the imputation methods and this could be explored in future research.

We now apply nearest neighbour imputation, hot deck imputation within classes and propensity score weighting to LFS data. All methods are weighted by the survey weights. Figure 1 compares an estimated distribution, which ignores measurement error (the bold line) with estimates based on three missing data methods (the three dotted lines). We suggest the latter estimates are more approximately unbiased than the former estimate. All three missing data adjustments show, as expected, a strong 'kink' in the distribution at the level of the national minimum wage unlike for the derived variable. Corresponding estimates of two low pay proportions of interest are presented in Table 6. The 'missing data adjustments' have a substantial impact in comparison to estimates based on the derived variable. The results suggest that the proportion of jobs paid at or below the national minimum wage rate may be overestimated by four or five times if measurement error is ignored. The differences between the missing data methods are much smaller. We can see that the estimates under propensity score weighting differ from estimates derived using imputation methods, at least for the June–August 1999 quarter. Note that this quarter of the LFS was subject to a lower response rate than subsequent quarters resulting from changes in the LFS questionnaire. It was found that for consecutive quarters, which are subject to about 43% response rate, weighting and imputation led to very similar estimates of low pay proportions, as illustrated in table 7 for the March–May 2000 quarter. The decrease in the proportion of low paid employees over time is a result of the impact of the National Minimum Wage legislation. In addition, different imputation and propensity score models are used to analyse the effects of various model specifications on estimates of low pay. From Table 6 we can see that there is an indication that different models can have an effect on the estimates. With increasing complexity of the model a reduction in the estimates for both point estimators is observed. This might reflect a departure from the MAR assumption for the simpler imputation models. At least for the 1999 quarter, the differences in the estimates between weighting and imputation methods seem to be greater than between models. Note that the estimates presented here might differ slightly from official UK estimates since, for example, the official estimates are based on different imputation models, treating outliers differently or imputing differently for certain professions.

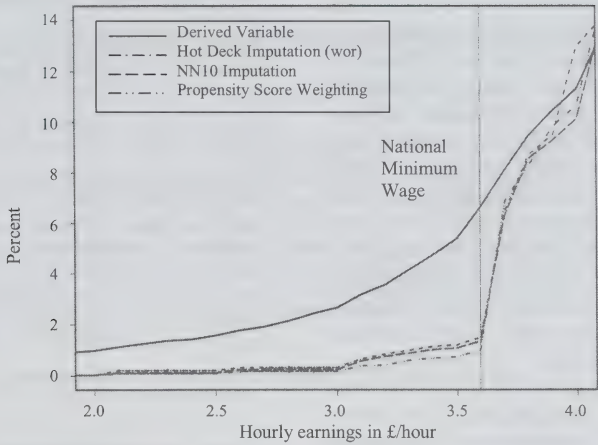


Figure 1. Alternative Estimates of the Distribution of Hourly Earnings From £2 to £4 for Age Group 22+, June-August 1999.

Table 6
Estimates of θ_1 and θ_2 (Weighted) for 18+ Using Different Propensity Score Models and Imputation Models Applied to LFS, June–August 1999

Method	Propensity Score Model or Imputation Model	(Weighted) $\hat{\theta}_1$ (%)	(Weighted) $\hat{\theta}_2$ (%)
Derived Variable	–	7.13	20.5
Propensity Score Weighting	M1	0.96	34.5
	M2	1.08	38.4
	M3	1.08	38.4
HDIWOR10	M1	1.44	32.1
	M2	1.41	32.9
	M3	1.50	33.2
NN10	M1	1.32	32.6
	M2	1.44	32.8
	M3	1.50	33.0

Note: M1 is the most complex model including square terms and interactions
M2 excludes the interactions and the square terms from model M1
M3 drops further covariates from model M2.

Table 7
Estimates of θ_1 and θ_2 (Weighted) for 18+ Using Propensity Score Weighting and Imputation Applied to LFS, March–May 2000

Method	Propensity Score Model or Imputation Model	(Weighted) $\hat{\theta}_1$ (%)	(Weighted) $\hat{\theta}_2$ (%)
Propensity Score Weighting	M1	0.54	27.10
HDIWOR10	M1	0.57	26.01
NN10	M1	0.55	26.61

9. Conclusions

In this paper we have considered the application of alternative missing data methods to correct for bias in the estimation of a distribution function arising from measurement error. Among imputation methods, nearest neighbour methods have performed most promisingly in terms of bias. These deterministic methods display no evidence of greater bias than stochastic imputation methods. Fractional imputation has shown appreciable efficiency gains compared to single imputation and appears more effective than penalizing the distance function or sampling without replacement with single imputation. In comparison to a propensity score weighting approach, the fractional nearest neighbour imputation has performed similarly, but has demonstrated slight advantages of robustness and efficiency. The simulation study suggested that the impact on the bias under a wrong model is greater for propensity score weighting and that the standard errors for the weighting approach were approximately 5–15% times higher than for the imputation method.

Further research is being undertaken to develop and evaluate associated variance estimation methods, as well as alternative point estimation methods based upon the Common Measurement Error Model in section 2.

Acknowledgements

We are grateful to Danny Pfeffermann for comments on an earlier version of this paper.

References

- Brick, J.M., and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.
- Buonaccorsi, J.P. (1990). Double sampling for exact values in some multivariate measurement error problems. *Journal of the American Statistical Association*, 85, 1075-1082.
- Chen, J., and Shao, J. (2000). Nearest neighbour imputation for survey data. *Journal of Official Statistics*, 16, 113-131.
- Chen, J., and Shao, J. (2001). Jackknife variance estimation for nearest neighbour imputation. *Journal of the American Statistical Association*, 96, 453, 260-269.
- Chesher, A. (1991). The effect of measurement error. *Biometrika*, 78, 451-462.
- David, M.H., Little, R., Samuël, M. and Triest, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 168-173.
- Dickens, R., and Manning, A. (2004). Has the national minimum wage reduced UK wage inequality? *Journal of the Royal Statistical Society, Series A*, 4, 613-626.
- Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.
- Fay, R.E. (1999). Theory and application of nearest neighbour imputation in census 2000. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 112-121.
- Fuller, W.A. (1995). Estimation in the presence of measurement error. *International Statistical Review*, 63, 121-141.
- Kalton, G. (1983). *Compensating for missing survey data*. Michigan, Institute for Social Research.
- Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics, Part A, Theory and Methods*, 13, 1919-1939.
- Kim, J.K. (2004). Efficient nonresponse weighting adjustment using estimated response probability. *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- Kim, J.-K., and Fuller, W.A. (2002). Variance estimation for nearest neighbour imputation. Unpublished manuscript.
- Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-301.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical analysis with missing data*. New York: John Wiley & Sons, Inc.
- Luo, M., Stokes, L. and Sager, T. (1998). Estimation of the CDF of a finite population in the presence of a calibration sample. *Environmental and Ecological Statistics*, 5, 277-289.
- Moore, J.C., Stinson, L.L. and Welniak, E.J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics*, 16, 331-361.
- ONS (1999). *Labour Force Survey*. User Guide, Volume 1, Background and Methodology, London.
- Rancourt, E. (1999). Estimation with nearest neighbour imputation at Statistics Canada. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 131-138.
- Rodgers, W.L., Brown, C. and Duncan, G.J. (1993). Errors in survey reports of earnings, hours worked and hourly wages. *Journal of the American Statistical Association*, 88, 1208-1218.
- Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.
- Rubin, D.B., and Schenker N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith), Chichester, Wiley.
- Skinner, C., Stuttard, N., Beissel-Durrant, G. and Jenkins, J. (2002). The measurement of low pay in the UK Labour Force Survey. *Oxford Bulletin of Economics and Statistics*, 64, 653-676.
- Stuttard, N., and Jenkins, J. (2001). Measuring low pay using the new earnings survey and the Labour Force Survey. *Labour Market Trends*, January 2001, 55-66.
- Tenenbein, A. (1970). A double sampling scheme for estimating from binary data with misclassifications. *Journal of the American Statistical Association*, 65, 1350-1361.

On Calibration Estimation for Quantiles

Torsten Harms and Pierre Duchesne¹

Abstract

In this paper, we consider the estimation of quantiles using the calibration paradigm. The proposed methodology relies on an approach similar to the one leading to the original calibration estimators of Deville and Särndal (1992). An appealing property of the new methodology is that it is not necessary to know the values of the auxiliary variables for all units in the population. It suffices instead to know the corresponding quantiles for the auxiliary variables. When the quadratic metric is adopted, an analytic representation of the calibration weights is obtained. In this situation, the weights are similar to those leading to the generalized regression (GREG) estimator. Variance estimation and construction of confidence intervals are discussed. In a small simulation study, a calibration estimator is compared to other popular estimators for quantiles that also make use of auxiliary information.

Key Words: Calibration estimators; Quantiles; Ratio estimators; Difference estimators.

1. Introduction

In recent years, considerable attention has been given to the estimation of population distribution functions in the context of survey sampling. A particular target of this attention has been the median, which is often regarded as a more satisfactory location measure than the mean, especially when the variable of interest follows a skewed distribution. Traditional estimators of population means or totals can be usually substantially improved if relevant auxiliary information is made available. Consequently, the use of such auxiliary information seems highly desirable in sample quantile estimators.

Using a model-based approach, Chambers and Dunstan (1986) considered quantile estimators based on an estimator of the distribution function which do incorporate auxiliary information. Rao, Kovar and Mantel (1990) have proposed design-based alternatives to the model-based approach. They used simulation experiments to compare two quantile estimators, based on ratio and difference estimators, to the simple design-based estimator which makes no use of the auxiliary information. It should be noted that neither of the two design-based proposals requires knowledge of the auxiliary information for each unit in the population; it rather suffices to know only the corresponding quantiles. While the model-based estimator proposed by Chambers and Dunstan (1986) can be more efficient than its design-based alternative if the model is correctly specified, Rao *et al.* (1990) have pointed out the advantage of the design-based estimators under model misspecification. Chambers, Dorfman and Hall (1992) have compared these two estimators theoretically with respect to their consistency, asymptotic bias and variance under a population model. Their main conclusion is that neither of the two methods is a

sharp winner. Dorfman (1993) has reevaluated the simulation results obtained by Rao *et al.* (1990) and proposed a modified version of their methodology, using model-based arguments. Variance estimators in the model-based approach of Chambers and Dunstan (1986) and the design-based estimators of Rao *et al.* (1990) are discussed in Wu and Sitter (2001).

Other related works on quantile and median estimators include that of Kuk (1988) who proposes quantile estimators under pps (*proportional to size*) sampling and that of Kuk and Mak (1989) who use a method that is based on cross-classifying the individuals in the sample, according to the variable of interest and a single auxiliary variable. Meeden (1995) takes a different approach to construct a median estimator based on univariate auxiliary information, using the Bayesian concept of Polya sampling to impute all the target variable's unknown population values via a ratio-based approach. Rueda, Arcos and Martínez (2003) have recently built quantile estimators that extend ratio, difference and regression estimators in ways similar to those developed for the population mean.

In this paper, we follow the concept of calibration which was first introduced by Deville (1988) in order to derive a quantile estimator. The calibration approach has gained popularity in real applications, because the resulting estimators are easy to interpret and to motivate, relying, as they do, on sampling weights and natural calibration constraints. This approach was developed in the seminal work of Deville and Särndal (1992) as an alternative means of incorporating auxiliary information in the estimation of population totals. The so-called calibrated weights are found by minimizing a distance measure between the sampling weights and the new weights, which need to satisfy certain calibration constraints. For estimating totals the calibrated weights replace

1. Torsten Harms and Pierre Duchesne, Université de Montréal, Département de mathématiques et de statistique, CP 6128 Succursale Centre-Ville, Montréal, Québec, H3C 3J7, Canada. E-mail: duchesne@dms.umontreal.ca.

the original design weights used in Horvitz-Thompson type estimators. When the new weights are applied to the auxiliary variables available in the sample, they reproduce the known population totals of the auxiliary variables exactly; it is for this reason that the estimators in this class are called calibration estimators. See also Singh and Mohl (1996) who provide simple justifications of calibration estimators. They also present a very general and unifying treatment of calibration methods whose weights satisfy certain range restrictions and benchmark constraints.

Our fundamental aim is to propose calibration estimators for quantiles which are as easy to implement and interpret as the calibration estimators for totals developed by Deville and Särndal (1992). When compared to the quantile estimators available in the literature, the new calibration estimators should also be competitive with respect to their bias, variance, and coverage rates of the confidence intervals. Early calibration estimators for distribution functions and quantiles include those proposed by Kovačević (1997), who considered estimators of the distribution function calibrated on moments of the auxiliary variables. Harms (2003) has investigated a similar approach, with applications to the Finnish European Household Panel survey. Ren (2002) appears to have been the first to develop a unifying treatment of calibration estimators for distribution functions and quantiles. The calibration estimators for quantiles presented in this paper continue the work initiated by Ren (2002). We adhere to the original calibration paradigm for totals as closely as possible: when the parameter of interest is a total, it seems natural to calibrate on totals of the auxiliary variables. In the present context, since the parameter of interest corresponds to a quantile, the calibration constraints require that the weights are such that the sample quantile estimators of the auxiliary variables and their corresponding population quantiles are equal. In other words, the weighted quantile estimators for the auxiliary variables should yield exactly the population quantiles, which are assumed to be known. We present arguments which justify calibrating on quantiles, whenever the parameter of interest is itself a quantile. Interestingly, our methodology does not necessitate knowledge of the values of the auxiliary variables for all units in the population. Since the resulting estimators display a structural form very similar to the original calibration estimators for totals, it is expected that, under general conditions, the proposed estimators for quantiles will be asymptotically design-unbiased. Furthermore, these similarities allow us to derive variance estimators which admit a familiar form. Contrary to some of the other estimators, the proposed approach is also applicable to vectorial auxiliary variables (that is, when several auxiliary variables are available), while requiring only minimal auxiliary information. However, some restrictions may apply when the

sample is highly unrepresentative of the sampled population or when the quantiles being estimated are very close to the population minimum or maximum. Note that highly unrepresentative samples can also cause problems for calibration estimators for totals commonly used; in such situations, the algorithm for computing calibration estimators may fail to converge for many distance measures of practical interest.

The organization of the paper is as follows: In section 2, some preliminaries are given, including a brief review of the calibration estimators for totals. The new calibration estimators for quantiles are developed in section 3.1. The standard distribution function can be interpreted as a Horvitz-Thompson estimator, providing a possible approach to the construction of a calibrated distribution function estimator. Quantile estimators are then naturally derived by inverting the distribution function estimator (see *e.g.*, Ren (2002)). As in calibration estimators for totals, design weights can be replaced by more general sampling weights, in order to take account the auxiliary information. However, for many situations of practical interest, it may happen that no solution exists for the calibration constraints when this kind of distribution function estimator is adopted, the reason being that this estimator corresponds to a step function. In order to avoid existence problems of solutions for the calibration constraints, a new distribution function estimator is introduced, based on the natural concept of interpolation. Under the common quadratic metric, an analytic representation of the calibration weights is provided in section 3.2; variance estimators and confidence intervals are discussed in section 3.3. A practical aspect involves evaluating the methodology proposed with real populations and several sampling plans. Consequently, in section 4, we present a small simulation study where we compare our new approach, with respect to variance, bias and coverage rates of the confidence intervals, with that of Chambers and Dunstan (1986) as well as with some of the estimators proposed by Rao *et al.* (1990). Finally, concluding remarks are offered in section 5.

2. Some Preliminaries on Calibration Estimators

In this section, we present the fundamental concepts and notations useful for the sequel. We also give a brief review of calibration estimators for totals.

Let $U = \{1, \dots, k, \dots, N\}$ be a finite population of size N . Let $T_y = \sum_{i \in U} y_k$ be the population total of the variable of interest y , (note that for a set A , $A \subseteq U$, \sum_A will be used as shorthand for $\sum_{k \in A}$). A sample $s \subset U$ of size n is drawn according to a sampling plan. Let $\pi_k = \Pr(s \ni k)$ and $\pi_{kl} = \Pr(s \ni k, l)$ be the first and second order inclusion probabilities, respectively. We denote the design

weights $d_k = \pi_k^{-1}$ and $\hat{T}_{y,HT} = \sum_s d_k y_k$ represents the Horvitz-Thompson (HT) estimator of T_y .

Let $\mathbf{x}_k = (x_{1k}, \dots, x_{Jk})'$ be a vector of auxiliary variables associated with unit k , $k \in U$. Calibration estimators naturally include auxiliary information in the estimation. Let $s = \{k_1, \dots, k_n\}$, $s \subset U$. Starting with the vector of original weights $\mathbf{d} = (d_{k_1}, \dots, d_{k_n})'$, new weights are found which, when applied to the auxiliary variables available in s , make it possible to retrieve the known population totals for the J auxiliary variables $\mathbf{T}_x = \sum_U \mathbf{x}_k = (T_{x_1}, \dots, T_{x_J})'$. The calibration estimator for totals are more precisely defined in Definition 1.

Definition 1 (Calibration estimator for totals). Let $\mathbf{d} = (d_{k_1}, \dots, d_{k_n})'$ be the design weights. The calibration estimator for totals takes the form $\hat{T}_{y,cal} = \sum_s w_{ks} y_k$, where the weights w_{ks} , $k \in s$ are obtained as the following minimization problem with respect to the variable $\mathbf{v} = (v_{k_1}, \dots, v_{k_n})'$:

$$\mathbf{w} = \operatorname{argmin}_{\mathbf{v}} D(\mathbf{v}, \mathbf{d}), \tag{1}$$

subject to the calibration constraints $\sum_s v_k \mathbf{x}_k = \mathbf{T}_x$, where $D(\cdot, \cdot)$ denotes the distance measure and $\mathbf{w} = (w_{k_1}, \dots, w_{k_n})'$ corresponds to the vector of the calibrated weights.

For notational simplicity, we write $w_k \equiv w_{ks}$ in Definition 1 when no confusion is possible. It is common practice to let $x_{1k} \equiv 1$, $\forall k \in U$, and consequently $T_{x_1} = N$. This means that the calibrated weights satisfy the natural constraint $\sum_s w_k = N$. Many distance functions D are available in the literature (see, e.g., Deville and Särndal (1992), Chen and Qin (1993), Thompson (1997)). Consider the quadratic distance function

$$D(\mathbf{v}, \mathbf{d}) = \sum_s \frac{(v_k - d_k)^2}{d_k q_k}, \tag{2}$$

where q_k determines the importance of the unit $k \in s$ in the calibration problem. Heteroscedasticity problems can be handled using an appropriate choice of the q_k 's. Solving the optimization problem (1) using the Lagrange multiplier technique (see Deville and Särndal (1992), among others), the weights $w_k = d_k (1 + q_k \mathbf{x}'_k \boldsymbol{\lambda}_s)$ are obtained, where $\boldsymbol{\lambda}_s = (\sum_s d_k q_k \mathbf{x}_k \mathbf{x}'_k)^{-1} (\mathbf{T}_x - \hat{\mathbf{T}}_{x,HT})$ and $\hat{\mathbf{T}}_{x,HT}$ denotes the HT-estimator of \mathbf{T}_x . This choice of distance function leads to the weights of the well-known generalized regression estimator (GREG) of Cassel, Särndal and Wretman (1976), which is studied in detail in Särndal, Swensson and Wretman (1992). Under minimal requirements for the distance measure D , Deville and Särndal (1992) have shown that all calibration estimators in this class are asymptotically equivalent to the GREG. For ease of interpretation and other cosmetic reasons, some users may want to have positive weights or restrict them to a specific interval (see also Singh

and Mohl (1996)). In practical applications, these numerical features of the weights seem to be the main motivation for an alternative choice of D .

3. New Calibration Estimators

In this section we develop calibration estimators for quantiles, using ideas similar to those leading to the calibration estimators for population totals, as described in section 2. The new calibration estimators for quantiles are introduced in the next subsection, using interpolated distribution function estimators. Then, special attention is devoted to the quadratic distance function. The last subsection presents variance estimation and the construction of confidence intervals.

3.1 Definition of the Calibration Estimators for Quantiles

Let $\mathbf{Q}_{x,\alpha} = (Q_{x_1,\alpha}, \dots, Q_{x_J,\alpha})'$ denote the known vector of population quantiles for the vector of auxiliary variables $\mathbf{x}_k = (x_{1k}, \dots, x_{Jk})'$, $k \in U$. The Heavyside function $H(z)$ is given by:

$$H(z) = \begin{cases} 1, & z \geq 0, \\ 0, & z < 0. \end{cases}$$

The population distribution function of a scalar auxiliary variable x is defined in the usual way as $F_x(t) = N^{-1} \sum_U H(t - x_k)$, and the population quantile $Q_{x,\alpha}$ is obtained by letting $Q_{x,\alpha} = \inf \{t | F_x(t) \geq \alpha\}$.

The vector $\mathbf{Q}_{x,\alpha}$ contains quantiles of the auxiliary variables, obtained from information in past surveys or from available administrative sources. For example, for skewed distributions which are rather common in business and economic surveys, it seems more natural to keep in the record files the population medians rather than population means; in this case it seems natural to assume the knowledge of $\mathbf{Q}_{x,0.5}$. This suggests that, using the same approach as the one leading to calibration for totals described in section 2, the proposed estimator for the population quantile $Q_{y,\alpha}$ of the variable of interest y , noted $\hat{Q}_{y,cal,\alpha}$, could be obtained by inverting a certain estimator of the distribution function (that we discuss below), subject to calibration constraints such as $\hat{Q}_{x_j,cal,\alpha} = Q_{x_j,\alpha}$, $j = 1, \dots, J$. Following the usual interpretation, if the calibrated weights allow us to retrieve the known population quantiles of the auxiliary variables then, under certain conditions, they should produce reasonable estimators for the quantile of the variable of interest y .

More precisely, the calibrated weights are obtained by solving the following optimization problem:

$$\mathbf{w} = \operatorname{argmin}_{\mathbf{v}} D(\mathbf{v}, \mathbf{d}), \tag{3}$$

subject to the calibration constraints $\sum_s v_k = N$ and $\hat{Q}_{x, \text{cal}, \alpha} = (\hat{Q}_{x_j, \text{cal}, \alpha}, \dots, \hat{Q}_{x_J, \text{cal}, \alpha})' = \mathbf{Q}_{x, \alpha}'$.

The estimators $\hat{Q}_{x, \text{cal}, \alpha}$ and $\hat{Q}_{y, \text{cal}, \alpha}$ rely on the vector of weights \mathbf{w} , stemming from the solution of the calibration problem (3). To calculate these estimators for quantiles, we need to construct w -weighted estimators of the distribution function for variables x and y . Based on the sampling weights \mathbf{d} , a natural estimator of the sampling distribution function is given by

$$\hat{F}_y(t) = \sum_s d_k H(t - y_k) / \sum_s d_k, \quad (4)$$

which provides a consistent estimator of $F_y(t)$. Similarly, $F_{x_j}(t)$ can be consistently estimated by $\hat{F}_{x_j}(t) = \sum_s d_k H(t - x_{jk}) / \sum_s d_k$, $j = 1, \dots, J$. A w -weighted distribution function estimator of $F_{x_j}(t)$ is given by

$$\hat{F}_{x_j, \text{cal}}(t) = \sum_s w_k H(t - x_{jk}) / \sum_s w_k. \quad (5)$$

A similar formula holds for $\hat{F}_{y, \text{cal}}(t)$. These w -weighted estimators are considered in Ren (2002). However, if one estimates $Q_{x_j, \alpha}$ by $\hat{Q}_{x_j, \alpha} = \inf \{t \mid \hat{F}_{x_j}(t) \geq \alpha\}$, or makes a similar estimation using a w -weighted version, then it is generally not possible to reach an exact solution of the calibration problem (3). Indeed, if the previous definition is used to estimate the quantiles by inverting the distribution function using the previous definitions, then the constraints in the optimization problem (3) will not, in general, be fulfilled unless the sample s contains precisely a unit k such that $x_{jk} = Q_{x_j, \alpha}$. When J is large, this problem can be more pronounced. Furthermore, even if the sample does contain such a value, it is sometimes not possible to obtain the weights needed to minimize the distance function, the reason being that under certain circumstances, the weights fulfilling the calibration constraints form an open set, whereas the optimal weights lie precisely on the border of this set. The following example illustrates this situation.

Example 1:

Consider a population U of size $N = 30$, such that the population median of x is $Q_{x, 0.5} = 2$. A sample s of size $n = 3$ is drawn, and suppose that $x_k = k$, $\forall k \in s = \{1, 2, 3\}$. For simplicity, the distance measure $D(\mathbf{v}, \mathbf{d}) = \sum_s (v_k - d_k)^2$ is adopted; it is supposed that the sampling weights are $(d_1, d_2, d_3) = (15, 9, 6)$. Based on (5), the calibration constraint is $\hat{Q}_{x, \text{cal}, 0.5} = \inf \{t \mid \hat{F}_{x, \text{cal}}(t) \geq 0.5\} = 2$, which implies that $\sum_s w_k H(2 - x_k) \geq 15$ and $\sum_s w_k H(1 - x_k) < 15$. Equivalently, $w_1 + w_2 \geq 15$ and $w_1 < 15$. Thus we have to choose w_1 of the form $w_1 = 15 - \epsilon$, for $\epsilon > 0$. In this case, since $w_1 + w_2 + w_3 = 30$, we have that $D(\mathbf{v}, \mathbf{d}) = \epsilon^2 + (w_2 - 9)^2 + (w_2 - 9 - \epsilon)^2$, leading to the optimal solution $(w_1, w_2, w_3) = (15 - \epsilon, 9 + \epsilon/2, 6 + \epsilon/2)$. Consequently, for these weights $D(\mathbf{v}, \mathbf{d}) = 3\epsilon^2/2$, which is obviously minimized when $\epsilon \rightarrow 0$. However, the limit

reduces to $\mathbf{w} = (w_1, w_2, w_3) = (15, 9, 6)$ with $D(\mathbf{w}, \mathbf{d}) = 0$, but based on these weights $\hat{Q}_{x, \text{cal}, 0.5} = 1 \neq Q_{x, 0.5} = 2$.

However, these difficulties can be naturally avoided by considering a smooth estimator of the distribution function. For simplicity, we consider here a distribution function estimator calculated using a linear interpolation (another possibility is discussed in section 5), which is precisely defined in Definition 2.

Definition 2 (Interpolated distribution function estimators).
Define

$$\hat{F}_{y, \text{cal}}(t) = \frac{\sum_s w_k H_{y, s}(t, y_k)}{\sum_s w_k}, \quad (6)$$

$$\hat{F}_{x_j, \text{cal}}(t) = \frac{\sum_s w_k H_{x_j, s}(t, x_{jk})}{\sum_s w_k}, \quad (7)$$

where the Heavyside function H in (4) and (5) is replaced by the slightly modified function

$$H_{y, s}(t, y_k) = \begin{cases} 1, & y_k \leq L_{y, s}(t), \\ \beta_{y, s}(t) & y_k = U_{y, s}(t), \\ 0, & y_k > U_{y, s}(t), \end{cases} \quad (8)$$

where $L_{y, s}(t) = \max \{ \{y_k, k \in s \mid y_k \leq t\} \cup \{-\infty\} \}$, $U_{y, s}(t) = \min \{ \{y_k, k \in s \mid y_k > t\} \cup \{\infty\} \}$ and $\beta_{y, s}(t) = \{t - L_{y, s}(t)\} / \{U_{y, s}(t) - L_{y, s}(t)\}$. The function $H_{x_j, s}(t, x_k)$ is defined similarly. The estimators (6) and (7), based on the functions $H_{y, s}(t, y_k)$ and $H_{x_j, s}(t, x_k)$, are called interpolated distribution function estimators of $F_y(t)$ and $F_{x_j}(t)$, respectively.

The various quantities in (8) have easy interpretations: $L_{y, s}$ and $U_{y, s}$ represent the lower and upper neighbors of t in the sampled values $y_k, k \in s$, and $\beta_{y, s}(t)$ denotes the linear interpolation coefficient between these two quantities. In particular, for all $t \in \{y_k, k \in s\}$ we have $H_{y, s}(t, y_k) = H(t - y_k)$. Consequently, the relations $\hat{F}_{y, \text{cal}}(t) = \hat{F}_{y, \text{cal}}(t)$ are satisfied for all $t \in \{y_k, k \in s\}$. For all the other values of t , $\hat{F}_{y, \text{cal}}(t)$ consists of a linear interpolation between these quantities. In the following example, Example 1 is revisited using the interpolated distribution function estimator (7).

Example 2:

In Example 1, using the interpolated version (7), the constraints are now $w_1 + w_2 + w_3 = 30$ and $(w_1 + w_2) / (w_1 + w_2 + w_3) = 0.5$. Consequently $w_3 = 15$, $w_1 + w_2 = 15$. Simple algebra shows that the optimal solution is $(w_1, w_2, w_3) = (10.5, 4.5, 15)$, which is now well-defined.

With the interpolated distribution function estimators, $\hat{F}_{y, \text{cal}}^{-1}(\alpha)$ and $\hat{F}_{x_j, \text{cal}}^{-1}(\alpha)$ are now well defined α -quantile estimators for all $\alpha \in (0, 1)$, as long as one can assure that the weights w_k are all strictly positive. Letting $\hat{Q}_{x_j, \text{cal}, \alpha} = \hat{F}_{x_j, \text{cal}}^{-1}(\alpha)$, we define the proposed calibration estimator

$\hat{Q}_{y, \text{cal}, \alpha}$ for the quantile $Q_{y, \alpha}$, using the interpolated distribution function estimator given in Definition 2.

Definition 3 (Calibration estimator for quantiles). *Consider the optimization problem (3), subject to the calibration constraints $\sum_s v_k = N$ and $\hat{\mathbf{Q}}_{x, \text{cal}, \alpha} = (\hat{Q}_{x, \text{cal}, \alpha}, \dots, \hat{Q}_{x_j, \text{cal}, \alpha})' = \mathbf{Q}_{x, \alpha}$. Solving this optimization problem and denoting the resulting weights as \mathbf{w} , the proposed calibration estimator for quantiles of $Q_{y, \alpha}$ is defined by*

$$\hat{Q}_{y, \text{cal}, \alpha} = \hat{F}_{y, \text{cal}}^{-1}(\alpha), \quad (9)$$

where $\hat{F}_{y, \text{cal}}(t)$ is given by (6).

One of the appealing properties of the proposed estimator (9) is that it yields exact population quantiles when the relationship between y and a scalar auxiliary variable x is exactly linear. Assume that $y_k = a + bx_k$ holds perfectly for all units $k \in U$ and suppose that the units in the sample s are such that $x_k < Q_{x, \alpha} < x_l$ for some units x_k and x_l , $k, l \in s$. For the calibrated estimator (9), we have that $\hat{F}_{x, \text{cal}}(Q_{x, \alpha}) = \alpha$. We need to distinguish the two cases, $b > 0$ and $b < 0$ (The case $b = 0$ is trivial since y_k is then identically equal to a constant). Firstly, consider the situation $b > 0$. Since the linear relation $y_k = a + bx_k$ is satisfied for all units k and since $b > 0$, the following relations hold: $L_{y, s}(a + bt) = a + bL_{x, s}(t)$; $U_{y, s}(a + bt) = a + bU_{x, s}(t)$ and $\beta_{y, s}(a + bt) = \beta_{x, s}(t)$. These relations lead to $H_{y, s}(a + bt, y_k) = H_{x, s}(t, x_k)$. It follows that $\hat{F}_{y, \text{cal}}(a + bt) = \hat{F}_{x, \text{cal}}(t)$. Furthermore, $\hat{F}_{y, \text{cal}}(a + bQ_{x, \alpha}) = \alpha$ and using the relation $a + bQ_{x, \alpha} = Q_{y, \alpha}$, we deduce that $\hat{F}_{y, \text{cal}}(Q_{y, \alpha}) = \alpha$. Consequently, when an exact linear relationship holds and $b > 0$, $\hat{Q}_{y, \text{cal}, \alpha} = \hat{F}_{y, \text{cal}}^{-1}(\alpha) = Q_{y, \alpha}$. Secondly, consider the case $b < 0$. We deduce in this case the following relations: $L_{y, s}(a + bt) = a + bU_{x, s}(t)$; $U_{y, s}(a + bt) = a + bL_{x, s}(t)$; $\beta_{y, s}(a + bt) = 1 - \beta_{x, s}(t)$ and $H_{y, s}(a + bt, y_k) = 1 - H_{x, s}(t, x_k)$. Since $b < 0$, the relationship between the quantiles of x and y is given by $a + bQ_{x, \alpha} = Q_{y, 1-\alpha}$. Then, we deduce that $\hat{F}_{y, \text{cal}}(Q_{y, 1-\alpha}) = \hat{F}_{y, \text{cal}}(a + bQ_{x, \alpha}) = 1 - \hat{F}_{x, \text{cal}}(Q_{x, \alpha}) = 1 - \alpha$. Thus, in this situation, $Q_{y, 1-\alpha}$ is estimated exactly by $\hat{Q}_{y, \text{cal}, 1-\alpha}$. This means that, when an exact relation holds, if $b > 0$ the proposed calibration estimator $\hat{Q}_{y, \text{cal}, \alpha}$ yields perfect estimators with zero bias and variance of $Q_{y, \alpha}$. On the other hand, if $b < 0$ and calibrating on $Q_{x, \alpha}$, $Q_{y, 1-\alpha}$ is estimated exactly by $\hat{Q}_{y, \text{cal}, 1-\alpha}$ (which makes sense because the perfect linear relationship between x and y is such that the slope parameter is negative).

Note that when $\hat{F}_{y, \text{cal}}$ and $\hat{F}_{x_j, \text{cal}}$ are invertible at points $Q_{y, \alpha}$ and $Q_{x_j, \alpha}$, the calibration constraints in (3) can be rewritten in terms of the distribution functions, that is the calibration constraints based on the quantiles are equivalent to $\hat{F}_{x_j, \text{cal}}(Q_{x_j, \alpha}) = \alpha$, $j = 1, \dots, J$. This means that the

original calibration problem can be alternatively written in terms of distribution functions with the above constraints.

A natural question arises as to the existence of a solution to the optimization problem (3). Even when formulated with the interpolated distribution functions, it is not always possible to find a solution to (3). For example, if $Q_{x_j, \alpha}$ is smaller or larger than all values x_{jk} in the sample s , then $\hat{F}_{x_j, \text{cal}}(Q_{x_j, \alpha})$ will equal zero or one regardless of the choice of the weights \mathbf{w} . Thus in these cases it may happen that the calibration constraints cannot be fulfilled. However, when the sample's behavior differs widely from that of the target population, one should keep a very critical eye on any adjustment, and this situation can be considered somewhat extreme. In practice, this rarely occurs unless α is chosen very close to zero or one. Note that it may be impossible to obtain a solution when the sample size n is small. In these situations, the sample minimum or maximum could serve as a possible estimator or we could resort to the simple design-based estimator of the distribution function.

The second potential problem is that some weights w_k might be negative. In this case $\hat{F}_{y, \text{cal}}$ is no longer bijective. This is not a problem as long as $\hat{F}_{y, \text{cal}}^{-1}(\alpha)$ is still uniquely determined. This problem can be avoided by restricting all the weights to be strictly positive, using an appropriate metric $D(\cdot, \cdot)$. This approach has been adopted by Kovačević (1997) (for more details on distance functions yielding positive weights, see also Deville and Särndal (1992) and Singh and Mohl (1996)).

Remark 1:

The proposed distribution functions estimators (6) and (7) rely on a linear interpolation. In a unified way, the population distribution function, which is a step function as well, could also be defined using a linear interpolation. In practice, the two definitions differ only slightly in behavior, if the population N is sufficiently large. However, it should be noted that if the population size N is relatively small, it might be worth using an interpolation to define distribution functions.

Remark 2:

In the optimization problem (3), we calibrated on a particular quantile. This approach could be extended by allowing to calibrate on a finite set of quantiles, if such information is available. More precisely, suppose that for an auxiliary variable x , the α_m -quantiles Q_{x, α_m} , $m = 1, \dots, M$ are known, where $M < n - 1$. In this case, we could consider the calibration constraints $\hat{F}_{x, \text{cal}}(Q_{x, \alpha_m}) = \alpha_m$, $m = 1, \dots, M$ and solve the optimization problem (3) with these additional calibration constraints. Naturally, this information yields a more complete description of the distribution of the auxiliary variables; so the efficiency of the calibration estimators is expected to be higher.

Remark 3:

The proposed calibration estimator (9) is obtained by calibrating on population quantiles. Another possibility has been considered by Ren (2002) who calibrated on population moments, up to order m , of the same distribution. More precisely, Ren (2002) has proposed calibration estimators for quantiles satisfying constraints of the form $\sum_s w_k x_k^m = \sum_U x_k^m$, $m = 0, 1, \dots, M$. Calibration on different moments of the same distribution is closely related to calibrating on different quantiles of the same variable, and all these constraints provide a more complete description of the distribution of the auxiliary variable. For other generalizations of the calibration paradigm on moments, see also Ren and Deville (2000) and Harms (2003).

3.2 Analytical Solution of the Calibrated Weights when D is the Quadratic Metric

When the quadratic distance function (2) is adopted, an explicit solution of the optimization problem (3) can be derived. This situation is similar to the calibration estimators for totals, where the weights of the GREG estimator are explicitly obtained under the metric (2). A careful analysis of the estimation problem for quantiles reveals important similarities, the reason being that the estimators given by (7) are weighted sums of the variables $\{H_{x_j, s}(t, x_{jk}), k \in s\}$, $j = 1, \dots, J$. This is stated in Proposition 1.

Proposition 1 (Calibrated weights for the quadratic metric). *Consider the quadratic distance function (2). The vector of weights \mathbf{w} which solves the optimization problem (3) satisfies the relation:*

$$w_k = d_k (1 + q_k \mathbf{a}'_k \boldsymbol{\lambda}_s), k \in s, \quad (10)$$

where the vector $\boldsymbol{\lambda}_s = (\lambda_0, \dots, \lambda_J)'$ is determined via the $J+1$ constraints as:

$$\boldsymbol{\lambda}_s = \left(\sum_s d_k q_k \mathbf{a}_k \mathbf{a}'_k \right)^{-1} (\mathbf{T}_a - \sum_s d_k \mathbf{a}_k), \quad (11)$$

with $\mathbf{T}_a = (N, \alpha, \dots, \alpha)'$ and the components of $\mathbf{a}_k = (1, a_{1k}, \dots, a_{Jk})'$ are given by

$$a_{jk} = \begin{cases} N^{-1}, & x_{jk} \leq L_{x_j, s}(Q_{x_j, \alpha}), \\ N^{-1} \beta_{x_j, s}(Q_{x_j, \alpha}), & x_{jk} = U_{x_j, s}(Q_{x_j, \alpha}), \\ 0, & x_{jk} > U_{x_j, s}(Q_{x_j, \alpha}), \end{cases}$$

with $j = 1, \dots, J$.

Proof. To prove Proposition 1, first note that, since the first constraint $\sum_s w_k = N$ must be satisfied, it follows that $\hat{F}_{x_j, \text{cal}}(t) = N^{-1} \sum_s w_k H_{x_j, s}(t, x_{jk})$. Proceeding as in Deville and Särndal (1992), we can show that the vector $\mathbf{a}_k = (1, a_{1k}, \dots, a_{Jk})'$ satisfies

$\mathbf{a}_k =$

$$\left(1, \frac{\partial \hat{F}_{x_1, \text{cal}}}{\partial w_k}, \dots, \frac{\partial \hat{F}_{x_J, \text{cal}}}{\partial w_k} \right)', \quad (12)$$

$\sum_s w_k = N; \hat{F}_{x_j, \text{cal}}(Q_{x_j, \alpha}) = \alpha, j = 1, \dots, J$

that we now evaluate explicitly. Evaluating the derivatives, we have that $a_{jk} = N^{-1} H_{x_j, s}(t, x_{jk})$, $j = 1, \dots, J$, evaluated at $t = Q_{x_j, \alpha}$. This leads to

$$a_{jk} = \begin{cases} N^{-1}, & x_{jk} \leq L_{x_j, s}(Q_{x_j, \alpha}), \\ N^{-1} \beta_{x_j, s}(Q_{x_j, \alpha}), & x_{jk} = U_{x_j, s}(Q_{x_j, \alpha}), \\ 0, & x_{jk} > U_{x_j, s}(Q_{x_j, \alpha}), \end{cases}$$

$j = 1, \dots, J$, as announced.

In (11), \mathbf{T}_a can be interpreted as the expected value of $\sum_s d_k \mathbf{a}_k$. The derived weights (10) in the distribution function estimator (6) rely on the variables \mathbf{a}_k , $k \in s$ defined by (12). Note that they correspond to a certain transformation of the auxiliary variable \mathbf{x}_k . The difference between the weights for totals and quantiles relies on this variable \mathbf{a}_k : when \mathbf{a}_k is replaced by \mathbf{x}_k , we retrieve the original weights for totals. Consequently, it is useful to interpret this new variable. When estimating a total, the impact on the j^{th} calibration constraint is measured by x_{jk} , for each unit $k \in s$. In our framework, the impact of the unit k is now given by N^{-1} if $x_{jk} \leq L_{x_j, s}(Q_{x_j, \alpha})$; it corresponds to the factor $N^{-1} \beta_{x_j, s}(Q_{x_j, \alpha})$ when $x_{jk} = U_{x_j, s}(Q_{x_j, \alpha})$ and it is null elsewhere. In section 5, we shall discuss other estimation problems, leading to different variables \mathbf{a}_k .

Noting the similarities between the estimation of totals and quantiles, variance estimation can also be considered. This issue is addressed in the next subsection.

3.3 Variance Estimation and Confidence Intervals

As described in the previous section, the estimator $\hat{Q}_{y, \text{cal}, \alpha}$ displays several similarities to the usual GREG estimator for population totals. The transformed variables given by (12) provide the main difference between the calibration estimators for quantiles and totals. Interestingly, because of the structural similarity with the original calibration estimators, it is straightforward to derive a confidence interval for the proposed estimator $\hat{Q}_{y, \text{cal}, \alpha}$. We consider the construction of confidence intervals following Woodruff's (1952) approach. The confidence interval is given in Result 1.

Result 1 (Woodruff confidence interval for the calibration estimator for quantiles). *The confidence interval based on Woodruff's (1952) approach, using the calibration estimator (9) for the quantile $Q_{y, \alpha}$ is given by*

$$[\hat{F}_{y, \text{cal}}^{-1}(\hat{c}_{1y}), \hat{F}_{y, \text{cal}}^{-1}(\hat{c}_{2y})], \quad (13)$$

where $\hat{c}_{1y} = \alpha - z_{1-\gamma/2}[\hat{V}\{\hat{F}_{y, \text{cal}}(Q_{y, \alpha})\}]^{1/2}$ and $\hat{c}_{2y} = \alpha + z_{1-\gamma/2}[\hat{V}\{\hat{F}_{y, \text{cal}}(Q_{y, \alpha})\}]^{1/2}$. The resulting procedure yields an approximate confidence interval for $Q_{y, \alpha}$ at a specified $1 - \gamma$ confidence level.

Proof. Assuming that $\hat{F}_{y, \text{cal}, \alpha}(Q_{y, \alpha})$ is approximately normally distributed, it follows that $\Pr(c_{1y} \leq \hat{F}_{y, \text{cal}, \alpha}(Q_{y, \alpha}) \leq c_{2y})$ should approximately be equal to $1 - \gamma$, if one chooses

$$c_{1y} = \alpha - z_{1-\gamma/2}[V\{\hat{F}_{y, \text{cal}}(Q_{y, \alpha})\}]^{1/2}, \quad (14)$$

$$c_{2y} = \alpha + z_{1-\gamma/2}[V\{\hat{F}_{y, \text{cal}}(Q_{y, \alpha})\}]^{1/2}, \quad (15)$$

where z_γ denotes the γ^{th} - quantile of the $N(0, 1)$ standard normal distribution. Since $\hat{F}_{y, \text{cal}, \alpha}(Q_{y, \alpha})$ represents essentially a sample mean, a possible variance estimator justified by the classical Taylor linearization is given by

$$\hat{V}\{\hat{F}_{y, \text{cal}}(Q_{y, \alpha})\} = N^{-2} \sum_s \sum_k \frac{\Delta_{kl}}{\pi_{kl}} (w_k e_k) (w_l e_l), \quad (16)$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$; the weights $w_k, k \in s$, correspond to the calibrated weights (3) which reduce to (10) when D is the quadratic distance function (2); the residuals are given by $e_k = H_{y, s}(\hat{Q}_{y, \text{cal}, \alpha}, y_k) - \mathbf{a}_k' \hat{\mathbf{B}}_s$ where

$$\hat{\mathbf{B}}_s = \left(\sum_s w_k q_k \mathbf{a}_k \mathbf{a}_k' \right)^{-1} \sum_s w_k q_k \mathbf{a}_k H_{y, s}(Q_{y, \text{cal}, \alpha}, y_k)$$

represents the regression coefficient estimator. Since the constants c_{1y} and c_{2y} given by (14) and (15) rely on $V\{\hat{F}_{y, \text{cal}}(Q_{y, \alpha})\}$, we can estimate these quantities using the variance estimator (16).

In Result 1, note that Deville and Särndal (1992) advocated a w -weighted variance estimator similar to (16) for estimating the variance of the calibration estimators of the population totals. The performance of the proposed calibration estimator (9) and the confidence interval given by (13) are studied empirically in section 4.

4. Simulation Results

From a practical point of view, it is natural to inquire about the finite sample properties of the new calibration estimators and to compare them to popular estimators for quantiles available in the literature. In this section, simulation experiments are undertaken, to illustrate empirically the new estimators. In particular, their empirical bias and variance in real populations are investigated. The coverage properties of the confidence intervals represent another question of practical interest, which is also studied.

In partial answer to these questions, we carried out three small simulation studies. For several sampling plans and for real populations, the proposed calibration estimator for

quantiles is compared to its popular competitors. In the next subsection 4.1, we describe in detail the populations investigated and we discuss the sampling plans chosen. In subsection 4.2, the estimators included in the empirical study are presented and, in subsection 4.3, the frequentist measures (empirical bias, variance and mean squared error, coverage rates of the confidence intervals) are described. Our empirical results are analyzed in subsection 4.4.

4.1 Description of the Real Populations and the Sampling Plans

The real populations are displayed in Figures 1 to 6. The first population, noted MU284, is taken from Särndal *et al.* (1992, Appendix B). This population consists of $N = 284$ municipalities in Sweden. We retain as variable of interest the population in 1985 (variable P85), and we assume that the auxiliary information available is the population in 1975 (variable P75). Both variables are measured in thousands. In Figure 1, the variable P85 is expressed as a function of P75; as expected, the relationship between P85 and P75 is strongly linear. The variable P85 follows a highly skewed distribution, as shown in Figure 2. In this population, 500 samples were drawn according to simple random sampling without replacement (SRS). In addition, the same study was carried out under a sampling plan with unequal probabilities, the Poisson (PO) sampling scheme. The properties of the PO sampling plan are described in Särndal *et al.* (1992). Due to the wide range of values for y , it was not possible to construct sample selection probabilities π_k of the form $\pi_k \propto y_k$, since this would mean that some π_k had to be greater than one. For the purpose of our illustration, we determined selection probabilities using the relation $\pi_k \propto 0.2y_k + 0.05$ (we recognize that these π_k 's are idealized, since y_k is not available in practice). Under the SRS sampling plan (PO sampling plan), we considered the sample sizes (expected sample sizes) $n = 25$ and $n = 50$.

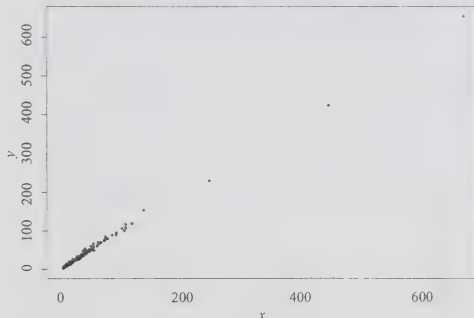


Figure 1. The Population MU284, where $y = \text{P85}$ and $x = \text{P75}$.

For the second study, we chose the MU284 population, but now made the variable of interest $y = \text{RMT85}$, which

represents the revenues from 1985 municipal taxation (in millions of kronor). Here the auxiliary variable chosen is $x = \text{REV84}$, which denotes real estate values according to 1984 assessments for each municipality (in millions of kronor). As can be seen in Figure 3, the relationship between x and y is somewhat spread out for larger values of x . The histogram of the variable RMT85 reveals that it follows a skewed distribution (Figure 4). For this study, 500 samples were drawn according to the SRS scheme of size $n = 25$ and $n = 50$.

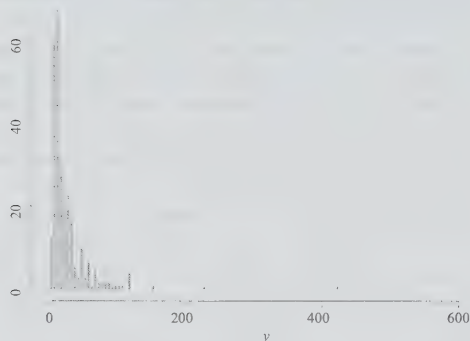


Figure 2. Histogram of the Variable P85 in the MU284 Population.

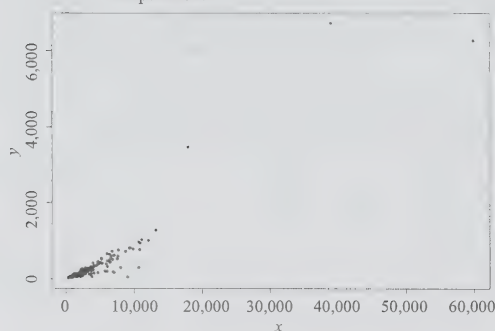


Figure 3. The Population MU284, where $y = \text{RMT85}$ and $x = \text{REV84}$.



Figure 4. Histogram of the Variable RMT85 in the MU284 Population.

The third population is based on a random subsample of the *Survey of Labor and Income Dynamics*, noted SLID982. The survey was conducted at Statistics Canada in 1998. For simplicity's sake, only entries with no missing values were selected. The size of the subsample is $N = 2,000$ and for our purpose this is assumed to be a population (the original sample size of this survey is approximately 60,000). Taxable income (in thousands of dollars) is the target variable and the auxiliary variable is the duration in months of the current employment. From Figure 5, the linear relationship between taxable income and length of employment is less pronounced. However, the two variables do not appear to be independent. In Figure 6, the variable of interest exhibits a strong coefficient of skewness. We have drawn 500 samples from the SLID982 population, according to SRS and PO sampling plans. The sample sizes (expected sample size) $n = 100$ and $n = 200$ were considered. For PO sampling, the first order probabilities, $\pi_k, k \in U$, were defined according to two rules. Under the first rule, the π_k 's were created such that π_k is approximately proportional to the variable of interest, that is taxable income (for the purpose of our study we assume that it is possible to create such π_k 's). Since some y_k are negative in this population, we chose $p_{1k} = y_k - \min\{y_k, k \in U\} + 1$ and we defined $\pi_k = E(n_s)p_{1k} / \sum_U p_{1k}$, where $E(n_s)$ stands for the expected sample size, in our case $E(n_s) = 100$ and 200. Under the second rule, the π_k 's were proportional to the entries in Table 1. This means that for each $k \in U$, there exists a factor p_{2k} , which is determined by the age-sex group of individual k . Then $\pi_k = E(n_s)p_{2k} / \sum_U p_{2k}$, where the factors p_{2k} are given in Table 1. The factors p_{2k} in Table 1 are based on a hypothetical sampling plan, in which we assume that these factors provide suitable size measures for the units in the various age-sex classes (see e.g., Särndal *et al.* (1992, page 87)); for these units, more males than females are likely to be selected and, for both sexes, adults in the 27 to 37 and 38 to 46 age range are more likely to be included in the sample.

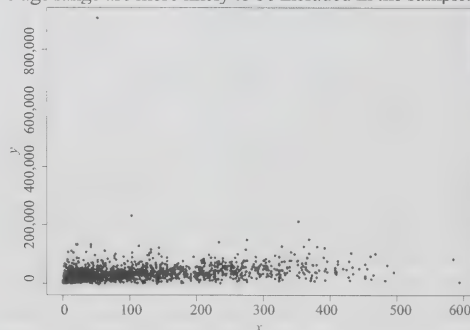


Figure 5. The SLID982 Population, where the Dependent Variable is the Taxable Income and Independent Variable is the Duration of Current Employment (in Months).

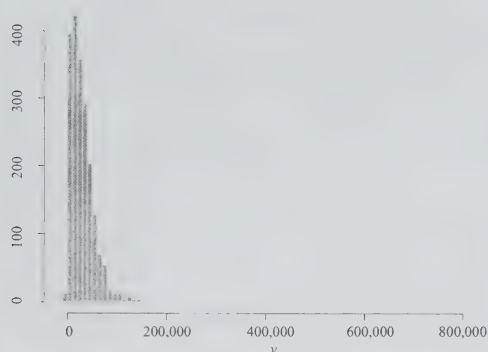


Figure 6. Histogram of the Taxable Income in the SLID982 Population.

Table 1
Factor p_{2k} by Age and Sex of Individual k ,
in the SLID982 Population

		Age			
		16–25	27–37	38–46	47–69
Sex	Male	3	6	5	4
	Female	1	2	3	2

In these three studies, we estimate the quartiles, that is the population parameters $Q_{y,\alpha}$ with $\alpha = 0.25, 0.5$ and 0.75 . Since the variables of interest display highly skewed distributions, it might be particularly interesting to study the quantile corresponding to $\alpha = 0.75$, in addition to the median and the first quartile. The next section describes the estimators included in the study.

4.2 Estimators Included in the Empirical Study

Since one of our goals is to propose estimators with reasonable properties with respect to bias, variance and coverage rates of the confidence intervals, we compare the new estimator defined by (9) based on the metric (2) to some of the popular quantile estimators proposed in the literature.

First, we include the simple design-based estimator based on the inversion of the estimator $\hat{F}_y(t) = \sum_s d_k H_{y,s}(t, y_k) / \sum_s d_k$:

$$\hat{Q}_{y,HT,\alpha} = \hat{F}_y^{-1}(\alpha). \quad (17)$$

The estimator (17) does not make use of auxiliary information. A possible variance estimator is

$$\hat{V}\{\hat{F}_y(Q_{y,\alpha})\} = \hat{N}^{-2} \sum \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \left\{ \frac{H_{y,s}(\hat{Q}_{y,HT,\alpha}, y_k) - \alpha}{\pi_k} \right\} \left\{ \frac{H_{y,s}(\hat{Q}_{y,HT,\alpha}, y_l) - \alpha}{\pi_l} \right\},$$

where $\hat{N} = \sum_s d_k$, and confidence intervals can be calculated using

$$[\hat{F}_y^{-1}(\tilde{c}_{1y}), \hat{F}_y^{-1}(\tilde{c}_{2y})],$$

where

$$\tilde{c}_{1y} = \alpha - z_{1-\gamma/2} [\hat{V}\{\hat{F}_y(Q_{y,\alpha})\}]^{1/2}, \quad (18)$$

$$\tilde{c}_{2y} = \alpha + z_{1-\gamma/2} [\hat{V}\{\hat{F}_y(Q_{y,\alpha})\}]^{1/2}. \quad (19)$$

For more details, see Särndal *et al.* (1992, page 202).

We also include in our empirical study the model-based estimator of Chambers and Dunstan (1986), which is motivated by a linear superpopulation model $y_k = \beta_0 + \beta'x_k + e_k$, $k \in U$, where e_k forms an identically and independently distributed sequence of random variables with mean zero and finite variance. Their estimator is defined as

$$\hat{Q}_{y,CD,\alpha} = \inf\{t \mid \hat{F}_{y,CD}(t) \geq \alpha\}, \quad (20)$$

where $\hat{F}_{y,CD}(t) = N^{-1} \{\sum_s H(t - y_k) + \sum_{U/s} \hat{G}(t - \hat{y}_k)\}$ represents a model-based estimator of the distribution function,

$$\hat{G}(u) = n^{-1} \sum_s H(u - \hat{e}_k) \quad (21)$$

denotes the empirical distribution function of the residuals $\hat{e}_k = y_k - \hat{y}_k$, $k \in s$, and $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}'x_k$, $k \in U/s$ correspond to the least-squares predictions. Since the estimator (20) basically imputes the unknown y_k for $k \in U/s$, note that it necessitates a complete knowledge of x_k for $k \in U$.

The construction of a confidence interval for $\hat{Q}_{y,CD,\alpha}$ relies on estimating the variance $V\{\hat{F}_{y,CD}(t)\}$. However, this variance estimation problem creates difficulties, since any analytical variance formula depends on the assumed model. Furthermore, such analytical expressions involve kernel density estimators, which are numerically intensive and depend on a kernel function and a bandwidth. For all these reasons, we decide to implement the delete-one jackknife variance estimators studied in Wu and Sitter (2001), who have shown the consistency of the proposed variance estimators. In the context of survey sampling, various resampling methods, including the jackknife, are introduced in Kovar, Rao and Wu (1988). The jackknife technique involves deleting a unit and re-calculating the estimator. Let $s_i = s \setminus \{i\}$ be the sample without unit i . Consider $\hat{\beta}_{0i}$ and $\hat{\beta}_i$, the regression estimators of β_0 and β calculated on s_i . Under a simple regression model, define

$$F_i^* = (n-1)^{-1} \sum_{k \in s_i} \left[\frac{N^{-1} \sum_{l \in U/s_i} H\{\hat{Q}_{y,CD,\alpha} - \hat{\beta}_i(x_l - x_k) - y_k\}}{H\{\hat{Q}_{y,CD,\alpha} - \hat{\beta}_i(x_i - x_k) - y_k\}} \right].$$

A consistent variance estimator of $V\{\hat{F}_{y,CD}(Q_{y,CD,\alpha})\}$ is given by

$$\begin{aligned}\hat{V}_{y, \text{CD}}\{\hat{F}_{y, \text{CD}}(Q_{y, \alpha})\} &= \frac{n-1}{n} \sum_{i \in s} (F_i^* - \bar{F}^*)^2 \\ &+ \frac{f(1-f)}{N-n} \sum_{k \in U/s} \{1 - \hat{G}(\hat{Q}_{y, \text{CD}, \alpha} - \hat{y}_k)\} \{1 - \hat{G}(\hat{Q}_{y, \text{CD}, \alpha} - \hat{y}_k)\},\end{aligned}$$

where $f = n/N$ is the sampling fraction, $\bar{F}^* = n^{-1} \sum_s F_i^*$, and \hat{G} is given by (21). Based on $\hat{V}_{y, \text{CD}}(Q_{y, \alpha})$, it is now possible to calculate the confidence intervals for $Q_{y, \alpha}$ using the inversion approach.

Since our method necessitates only the knowledge of the vector of quantiles $Q_{x, \alpha}$, we include in our study the ratio and difference estimators for the quantiles studied in Rao *et al.* (1990):

$$\hat{Q}_{y, \text{ra}, \alpha} = Q_{x, \alpha} (\hat{Q}_{y, \text{HT}, \alpha} / \hat{Q}_{x, \text{HT}, \alpha}), \quad (22)$$

$$\hat{Q}_{y, \text{diff}, \alpha} = \hat{Q}_{y, \text{HT}, \alpha} + \hat{R}(Q_{x, \alpha} - \hat{Q}_{x, \text{HT}, \alpha}), \quad (23)$$

where $\hat{Q}_{y, \text{HT}, \alpha}$ is given by (17) and $\hat{Q}_{x, \text{HT}, \alpha}$ is calculated similarly; the ratio estimator given by $\hat{R} = \sum_s d_k y_k / \sum_s d_k x_k$ provides a consistent estimator of $R = \sum_U y_k / \sum_U x_k$. Note that the estimators (22) and (23) are elaborated based on a scalar auxiliary variable, that is $J=1$. Valid variance estimators of (22) and (23) are given by:

$$\begin{aligned}\hat{V}(\hat{Q}_{y, \text{ra}, \alpha}) &= \hat{V}(\hat{Q}_{y, \text{HT}, \alpha}) \\ &+ \left(\frac{\hat{Q}_{y, \text{HT}, \alpha}}{\hat{Q}_{x, \text{HT}, \alpha}} \right)^2 \hat{V}(\hat{Q}_{x, \text{HT}, \alpha}) \\ &- 2 \frac{\hat{Q}_{y, \text{HT}, \alpha}}{\hat{Q}_{x, \text{HT}, \alpha}} \hat{C}(\hat{Q}_{y, \text{HT}, \alpha}, \hat{Q}_{x, \text{HT}, \alpha}), \\ \hat{V}(\hat{Q}_{y, \text{diff}, \alpha}) &= \hat{V}(\hat{Q}_{y, \text{HT}, \alpha}) \\ &+ \hat{R}^2 \hat{V}(\hat{Q}_{x, \text{HT}, \alpha}) \\ &- 2\hat{R} \hat{C}(\hat{Q}_{y, \text{HT}, \alpha}, \hat{Q}_{x, \text{HT}, \alpha}).\end{aligned}$$

These variance estimators rely on the variance of $\hat{Q}_{y, \text{HT}, \alpha}$, and the covariance between $\hat{Q}_{y, \text{HT}, \alpha}$ and $\hat{Q}_{x, \text{HT}, \alpha}$ which are estimated using Woodruff's (1952) approach:

$$\hat{V}(\hat{Q}_{y, \text{HT}, \alpha}) = \frac{W_y^2}{4z_{1-\gamma/2}^2},$$

$$\hat{C}(\hat{Q}_{y, \text{HT}, \alpha}, \hat{Q}_{x, \text{HT}, \alpha}) = \frac{W_y W_x \hat{C}\{\hat{F}_x(Q_{x, \alpha}), \hat{F}_y(Q_{y, \alpha})\}}{4z_{1-\gamma/2}^2 [\hat{V}\{\hat{F}_x(Q_{x, \alpha})\}]^{1/2} [\hat{V}\{\hat{F}_y(Q_{y, \alpha})\}]^{1/2}},$$

where $W_y = \hat{F}_y^{-1}(\tilde{c}_{2y}) - \hat{F}_y^{-1}(\tilde{c}_{1y})$ and $W_x = \hat{F}_x^{-1}(\tilde{c}_{2x}) - \hat{F}_x^{-1}(\tilde{c}_{1x})$ denote the Woodruff intervals associated with y and x , with \tilde{c}_{1y} and \tilde{c}_{2y} defined by (18) and (19), $\tilde{c}_{1x} = \alpha - z_{1-\gamma/2} [\hat{V}\{\hat{F}_x(Q_{x, \alpha})\}]^{1/2}$, $\tilde{c}_{2x} = \alpha + z_{1-\gamma/2} [\hat{V}\{\hat{F}_x(Q_{x, \alpha})\}]^{1/2}$ and

$$\begin{aligned}\hat{C}\{\hat{F}_y(Q_{y, \alpha}), \hat{F}_x(Q_{x, \alpha})\} &= \\ \hat{N}^{-2} \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} &\left\{ \frac{H_{y, s}(\hat{Q}_{y, \text{HT}, \alpha}, y_k) - \alpha}{\pi_k} \right\} \\ &\left\{ \frac{H_{x, s}(\hat{Q}_{x, \text{HT}, \alpha}, x_l) - \alpha}{\pi_l} \right\}.\end{aligned}$$

Summarizing, we expect $\hat{Q}_{y, \text{CD}, \alpha}$ to perform well when the linear model describes the population adequately. This motivates the comparison of the new methodology with a model-based estimator. Furthermore, it seems of interest to evaluate $\hat{Q}_{y, \text{cal}, \alpha}$ and the leading design-based proposals, such as $\hat{Q}_{y, \text{diff}, \alpha}$ and $\hat{Q}_{y, \text{ra}, \alpha}$. The estimators $\hat{Q}_{y, \text{cal}, \alpha}$, $\hat{Q}_{y, \text{diff}, \alpha}$ and $\hat{Q}_{y, \text{ra}, \alpha}$ use $Q_{x, \alpha}$ only to improve the estimations and they take into account the sampling plan; these estimators are natural competitors. Note that the different estimators included in our study are elaborated under different assumptions on the dimension of the vector of the auxiliary variable \mathbf{x} , and on the availability of \mathbf{x}_k . Table 2 provides a comparison of the different estimators described in this section.

Table 2

Comparison of the Proposed Calibration Estimators and of Some Leading Estimators for Quantiles Proposed in the Literature, with Respect to the Dimension J of \mathbf{x} and the information requirement on \mathbf{x}

Estimator	Dimension of \mathbf{x}	Information requirements on \mathbf{x}
$\hat{Q}_{y, \text{HT}, \alpha}$	n.a.	none
$\hat{Q}_{y, \text{CD}, \alpha}$	$J \geq 1$	$\mathbf{x}_k, k \in U/s$
$\hat{Q}_{y, \text{ra}, \alpha}$	$J = 1$	$Q_{x, \alpha}$
$\hat{Q}_{y, \text{diff}, \alpha}$	$J = 1$	$Q_{x, \alpha}$
$\hat{Q}_{y, \text{cal}, \alpha}$	$J \geq 1$	$Q_{x, \alpha}$

4.3 Frequentist Measures

Our goal is to evaluate the estimators with respect to bias and variance. Other important considerations are the mean squared error (MSE) and the coverage rates of the confidence intervals.

Let $\hat{Q}_{y, \alpha}$ be an estimator of the population quantile $Q_{y, \alpha}$. Assume $\hat{Q}_{y, \alpha}^{(v)}$ is the estimator of the quantile calculated using the sample v , $v = 1, \dots, K$. The Monte Carlo mean E_{MC} , the Monte Carlo bias B_{MC} , and the Monte Carlo variance V_{MC} are given by the usual formulas, that is

$$E_{\text{MC}}(\hat{Q}_{y, \alpha}) = K^{-1} \sum_{v=1}^K \hat{Q}_{y, \alpha}^{(v)},$$

$$B_{\text{MC}} = E_{\text{MC}}(\hat{Q}_{y, \alpha}) - Q_{y, \alpha},$$

$$V_{\text{MC}}(\hat{Q}_{y, \alpha}) = K^{-1} \sum_{v=1}^K \{\hat{Q}_{y, \alpha}^{(v)} - E_{\text{MC}}(\hat{Q}_{y, \alpha})\}^2.$$

Our main criterion for determining efficiency is the Monte Carlo MSE, defined by $MSE_{MC} = K^{-1} \sum_{v=1}^K (\hat{Q}_{y,\alpha}^{(v)} - Q_{y,\alpha})^2$. The confidence intervals are calculated at the 95% confidence level, according to the procedures described in the previous sections. For an estimator $\hat{Q}_{y,\alpha}^{(v)}$ and its variance estimator $\hat{V}^{(v)}$, $v = 1, \dots, K$, the coverage rates at the 95% confidence level are calculated as

$$CR(\hat{Q}_{y,\alpha}) = K^{-1} \sum_{v=1}^K I \left(\left\{ \left[\hat{Q}_{y,\alpha}^{(v)} - 1.96 \sqrt{\hat{V}^{(v)}}, \hat{Q}_{y,\alpha}^{(v)} + 1.96 \sqrt{\hat{V}^{(v)}} \right] \right\} \right),$$

where $I(A)$ is the indicator function of the set A . The coverage rates are given below the column CR. We recall that we adopt $K = 500$ for all studies.

4.4 Discussion of the Empirical Results

The results are presented in Tables 3 to 8. We first discuss the results from Tables 3 to 4, when sampling the MU284 population with SRS and PO sampling plans. As can be seen, all the estimators display a similar behavior in both studies. The model-based estimator $\hat{Q}_{y,CD,\alpha}$ appears to be the most efficient among those analyzed when examining $\alpha = 0.75$ and is in general very efficient. This was expected, since the relationship between $x = P75$ and $y = P85$ is strongly linear and the model-based estimator assumes a simple regression model. However, for $\alpha = 0.25$ the differences in efficiency are less pronounced with respect to the other estimators based on auxiliary information. Among the estimators using only $Q_{x,\alpha}$ as information on the auxiliary variable, a rather similar performance is obtained. When the sample size is small, coverage rates usually deviate from the 95% nominal level. This is particularly true for the coverage rates of $\hat{Q}_{y,cal,\alpha}$, which are somewhat underestimated. However, some improvement is observed at $n = 50$, illustrating the consistency of the procedures studied. On the other hand, those of $\hat{Q}_{y,ra,\alpha}$ and $\hat{Q}_{y,diff,\alpha}$ are always one. This suggests that the variances are overestimated for these estimators. Due to an important component of bias in the MSE, the coverage rates of the model-based estimator sometimes deteriorate as the sample size increases. The best coverage rates are obtained by using the simple HT estimator, $\hat{Q}_{y,HT,\alpha}$, which is however less efficient than the other estimators.

Table 5 shows the result for the second population, which is the MU284 population but with $y = RMT85$ and $x = REV84$. Figure 3 seems to show a heteroscedasticity phenomenon in this population. In view of this, since the ratio estimator is justified when the underlying population displays such behavior, it is not surprising that the ratio

estimator $\hat{Q}_{y,ra,\alpha}$ performs well in this particular situation; if outperforms $\hat{Q}_{y,diff,\alpha}$ in several cases. For a small sample size, the ratio estimator generally behaves better than $\hat{Q}_{y,cal,\alpha}$. However, for $n = 50$, the calibration estimator appears to perform as well or slightly better than the ratio estimator. In this experiment, the bias and variance of the model-based estimator $\hat{Q}_{y,CD,\alpha}$ increase the MSE substantially. Furthermore, in some cases, confidence intervals for this estimator could not be obtained, since the Woodruff method is not appropriate in cases with extremely large variance (the Woodruff interval becomes too large and the linearity of the distribution function within this interval can thus no longer be assumed). We suspect that a model taking into account heteroscedasticity might improve the performance of the model-based estimator. This highlights the fact that to obtain high efficiency with model-based estimators, the model must be correctly specified.

The results in Table 6 to 8 concern the SLID982 population, under SRS and PO sampling plans with two rules for the π_k 's. All the estimators in Table 6 perform reasonably well in estimating the first quartile and the median, except for the ratio estimator $\hat{Q}_{y,ra,\alpha}$ which is the least efficient. Since the relationship between the dependent and independent variables is not precisely a linear model, this may partially explain the poor performance of the ratio estimator in this case. The relationship between x and y is not proportional and so the difference estimator $\hat{Q}_{y,diff,\alpha}$ appears preferable to $\hat{Q}_{y,ra,\alpha}$. However, for $\alpha = 0.75$, these estimators show the highest MSE, being both the least efficient. Interestingly, in this part of the experiment $\hat{Q}_{y,cal,\alpha}$ dominates the design-based estimators in terms of MSE. However, for small α , $\hat{Q}_{y,diff,\alpha}$ and $\hat{Q}_{y,cal,\alpha}$ perform similarly. It should be noted that for a larger sample size, $\hat{Q}_{y,cal,\alpha}$ and $\hat{Q}_{y,CD,\alpha}$ give the best efficiencies for the median and the third quartile. In fact, the model-based estimator $\hat{Q}_{y,CD,\alpha}$ slightly outperforms $\hat{Q}_{y,cal,\alpha}$, but it should be noted that it uses more auxiliary information than $\hat{Q}_{y,cal,\alpha}$.

Tables 7 and 8 present results under PO sampling plans. In general, design-based estimators perform much like those under SRS sampling plan. This is not the case for the model-based estimator; it is less efficient, likely because it does not incorporate the information about the sampling plan. More precisely, Table 7 presents simulation results under PO sampling, using the first rule for the π_k 's, $k \in U$. Coverage rates of the model-based estimator are particularly disappointing in this experiment; the components of bias were too important in the MSE. The design-based estimators provide much closer empirical coverage rates, to the nominal 95% confidence level. For moderate and large α , $\hat{Q}_{y,cal,\alpha}$ is the most efficient estimator. In fact, the calibration estimator $\hat{Q}_{y,cal,\alpha}$ performs well in this

experiment. Finally, Table 8 presents results obtained under PO sampling with the second rule for the π_k 's. In this case, $\hat{Q}_{y,ra,\alpha}$ is the least efficient estimator for the first quartile

and the median, and $\hat{Q}_{y,diff,\alpha}$ is the least efficient for $\alpha = 0.75$. In general, $\hat{Q}_{y,cal,\alpha}$ dominates the other estimators in this situation, offering the highest efficiency.

Table 3
Monte Carlo Simulation Results for Sampling from the MU284 Population, $y = P85$, $x = P75$, Under SRS Sampling Plan.
The Number of Replications is Set at $K = 500$

α	Estimator	$n = 25$				$n = 50$			
		B_{MC}	V_{MC}	MSE_{MC}	CR	B_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y,cal,\alpha}$	-0.0343	0.5075	0.5077	0.886	-0.0499	0.2437	0.2457	0.828
	$\hat{Q}_{y,HT,\alpha}$	-0.0266	2.3196	2.3157	0.952	0.0035	1.1087	1.1065	0.936
	$\hat{Q}_{y,ra,\alpha}$	-0.1444	0.3869	0.4070	1.000	-0.0774	0.1684	0.1741	1.000
	$\hat{Q}_{y,diff,\alpha}$	-0.1486	0.3901	0.4114	1.000	-0.0734	0.1723	0.1774	1.000
	$\hat{Q}_{y,CD,\alpha}$	0.4855	0.2791	0.5143	0.906	0.5485	0.1981	0.4985	0.824
0.5	$\hat{Q}_{y,cal,\alpha}$	-0.2762	1.6499	1.7229	0.918	-0.2835	0.9585	1.0370	0.944
	$\hat{Q}_{y,HT,\alpha}$	0.2605	12.5161	12.5589	0.922	-0.0064	5.8466	5.8349	0.916
	$\hat{Q}_{y,ra,\alpha}$	-0.2586	0.8828	0.9479	1.000	-0.4296	0.6701	0.8533	1.000
	$\hat{Q}_{y,diff,\alpha}$	-0.2775	0.9898	1.0648	1.000	-0.4331	0.7492	0.9352	1.000
	$\hat{Q}_{y,CD,\alpha}$	0.9431	0.4054	1.2940	0.866	0.9884	0.2410	1.2175	0.714
0.75	$\hat{Q}_{y,cal,\alpha}$	-0.6229	3.3241	3.7055	0.614	-0.3661	1.8107	1.9411	0.710
	$\hat{Q}_{y,HT,\alpha}$	-0.1414	53.1951	53.1088	0.948	-0.3692	18.8586	18.9572	0.964
	$\hat{Q}_{y,ra,\alpha}$	-0.7925	3.0021	3.6242	1.000	-1.0004	1.4594	2.4573	1.000
	$\hat{Q}_{y,diff,\alpha}$	-0.8230	3.4379	4.1083	1.000	-1.0396	1.5267	2.6044	1.000
	$\hat{Q}_{y,CD,\alpha}$	0.4343	0.5108	0.6984	0.954	0.4485	0.2618	0.4624	0.974

Table 4
Monte Carlo Simulation Results for Sampling from the MU284 Population, $y = P85$, $x = P75$, Under PO Sampling Plan.
The Number of Replications is Set at $K = 500$

α	Estimator	$n = 25$				$n = 50$			
		B_{MC}	V_{MC}	MSE_{MC}	CR	B_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y,cal,\alpha}$	-0.0441	0.4886	0.4896	0.888	-0.0169	0.2601	0.2599	0.828
	$\hat{Q}_{y,HT,\alpha}$	-0.1698	2.2825	2.3068	0.936	-0.0384	1.1828	1.1819	0.928
	$\hat{Q}_{y,ra,\alpha}$	-0.1509	0.3857	0.4076	1.000	-0.0913	0.2100	0.2179	1.000
	$\hat{Q}_{y,diff,\alpha}$	-0.1634	0.3821	0.4080	1.000	-0.0877	0.2149	0.2221	1.000
	$\hat{Q}_{y,CD,\alpha}$	0.6709	0.3310	0.7805	0.896	0.8792	0.1339	0.9066	0.554
0.5	$\hat{Q}_{y,cal,\alpha}$	-0.3610	1.4881	1.6155	0.920	-0.3236	0.8833	0.9863	0.936
	$\hat{Q}_{y,HT,\alpha}$	-0.0612	11.3969	11.3778	0.926	-0.2712	5.2672	5.3302	0.906
	$\hat{Q}_{y,ra,\alpha}$	-0.3735	1.0009	1.1385	1.000	-0.4130	0.5486	0.7181	1.000
	$\hat{Q}_{y,diff,\alpha}$	-0.3962	1.1271	1.2818	1.000	-0.4217	0.5962	0.7729	1.000
	$\hat{Q}_{y,CD,\alpha}$	1.1740	0.4947	1.8719	0.820	1.3297	0.2146	1.9822	0.474
0.75	$\hat{Q}_{y,cal,\alpha}$	-0.6420	2.6605	3.0674	0.608	-0.4476	1.6212	1.8183	0.708
	$\hat{Q}_{y,HT,\alpha}$	-0.6200	51.2934	51.5752	0.956	-0.6632	17.3625	17.7677	0.966
	$\hat{Q}_{y,ra,\alpha}$	-0.8686	2.8841	3.6329	1.000	-0.9683	1.6494	2.5837	1.000
	$\hat{Q}_{y,diff,\alpha}$	-0.9025	2.9826	3.7911	1.000	-1.0177	1.6340	2.6665	1.000
	$\hat{Q}_{y,CD,\alpha}$	0.4620	0.4501	0.6627	0.982	0.5388	0.2329	0.5228	0.980

Table 5
Monte Carlo Simulation Results for Sampling from the MU284 Population, $y = \text{RMT85}$, $x = \text{REV84}$, Under SRS Sampling Plan.
The Number of Replications is Set at $K = 500$

α	Estimator	$n = 25$				$n = 50$			
		B_{MC}	V_{MC}	MSE_{MC}	CR	B_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y, \text{cal}, \alpha}$	1.0161	51.5421	52.4714	0.892	0.6499	24.0662	24.4404	0.954
	$\hat{Q}_{y, \text{HT}, \alpha}$	0.3733	110.2572	110.1760	0.960	0.3383	47.2921	47.3120	0.962
	$\hat{Q}_{y, \text{ra}, \alpha}$	3.0025	65.4135	74.2979	0.998	2.3856	30.7284	36.3580	0.992
	$\hat{Q}_{y, \text{diff}, \alpha}$	2.5952	107.7891	114.3084	0.994	2.4083	55.6977	61.3862	0.986
	$\hat{Q}_{y, \text{CD}, \alpha}$	-16.5165	1661.0257	1930.4983	0.990	-17.3217	820.7447	1119.1443	0.960
0.5	$\hat{Q}_{y, \text{cal}, \alpha}$	-1.6219	215.0326	217.2330	0.870	-0.3419	118.2125	118.0930	0.922
	$\hat{Q}_{y, \text{HT}, \alpha}$	0.0075	763.6236	762.0964	0.910	-0.3977	331.2357	330.7314	0.914
	$\hat{Q}_{y, \text{ra}, \alpha}$	0.7712	212.8298	212.9988	0.996	-0.2810	136.4382	136.2443	0.996
	$\hat{Q}_{y, \text{diff}, \alpha}$	0.3415	283.6718	283.2210	0.998	-1.0104	201.3707	201.9889	0.998
	$\hat{Q}_{y, \text{CD}, \alpha}$	17.6124	190.0045	499.8199	n.a.	13.5037	100.2106	282.3611	0.566
0.75	$\hat{Q}_{y, \text{cal}, \alpha}$	-5.3477	1023.6924	1050.2431	0.826	-4.7339	443.0660	464.5896	0.926
	$\hat{Q}_{y, \text{HT}, \alpha}$	-4.6352	3526.8202	3541.2514	0.938	-5.8890	1242.4858	1274.6812	0.940
	$\hat{Q}_{y, \text{ra}, \alpha}$	-1.4390	980.5573	980.6669	0.994	-2.0070	555.5135	558.4305	1.000
	$\hat{Q}_{y, \text{diff}, \alpha}$	-5.3988	1464.7867	1491.0041	0.996	-3.9008	744.1604	757.8881	1.000
	$\hat{Q}_{y, \text{CD}, \alpha}$	49.3038	2753.8212	5179.1826	n.a.	49.4089	1488.9734	3927.2324	0.596

Table 6
Monte Carlo Simulation Results for Sampling from the SLID982 Population, Under SRS Sampling Plan.
The Number of Replications is Set at $K = 500$

α	Estimator	$n = 100$				$n = 200$			
		BR_{MC}	V_{MC}	MSE_{MC}	CR	BR_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y, \text{cal}, \alpha}$	0.1360	3.0390	3.0514	0.956	0.2331	1.6787	1.7297	0.934
	$\hat{Q}_{y, \text{HT}, \alpha}$	-0.0596	3.6099	3.6062	0.946	0.0499	1.9277	1.9263	0.918
	$\hat{Q}_{y, \text{ra}, \alpha}$	0.3067	6.8815	6.9618	0.970	0.0910	3.0743	3.0764	0.958
	$\hat{Q}_{y, \text{diff}, \alpha}$	-0.0504	2.9691	2.9657	0.980	0.0198	1.6139	1.6111	0.952
	$\hat{Q}_{y, \text{CD}, \alpha}$	1.1042	2.1180	3.3329	0.922	1.1392	1.2937	2.5888	0.826
0.5	$\hat{Q}_{y, \text{cal}, \alpha}$	-0.4034	6.3364	6.4865	0.966	-0.1402	2.9940	3.0076	0.940
	$\hat{Q}_{y, \text{HT}, \alpha}$	-0.4157	7.4589	7.6168	0.918	-0.1894	3.5865	3.6151	0.928
	$\hat{Q}_{y, \text{ra}, \alpha}$	0.7015	41.8314	42.2399	0.958	0.2238	18.7005	18.7131	0.952
	$\hat{Q}_{y, \text{diff}, \alpha}$	-0.4859	14.2083	14.4160	0.970	-0.2740	6.6184	6.6803	0.974
	$\hat{Q}_{y, \text{CD}, \alpha}$	0.5702	3.5420	3.8601	0.952	0.6697	1.7559	2.2009	0.932
0.75	$\hat{Q}_{y, \text{cal}, \alpha}$	-0.4164	12.4657	12.6142	0.952	-0.2384	5.9118	5.9568	0.950
	$\hat{Q}_{y, \text{HT}, \alpha}$	-0.5913	12.5456	12.8701	0.930	-0.3519	6.5496	6.6603	0.926
	$\hat{Q}_{y, \text{ra}, \alpha}$	0.7404	48.6836	49.1345	0.954	0.2967	18.5786	18.6294	0.966
	$\hat{Q}_{y, \text{diff}, \alpha}$	0.3288	53.6456	53.6464	0.954	0.1841	21.7552	21.7456	0.966
	$\hat{Q}_{y, \text{CD}, \alpha}$	0.5966	8.3416	8.6809	0.954	0.5413	4.3692	4.6535	0.936

Table 7

Monte Carlo Simulation Results for Sampling from the SLID982 Population, Under PO Sampling Plan and the First Rule for the Construction of the $\pi_k, k \in U$. The number of replications is set at $K = 500$

α	Estimator	$n = 100$				$n = 200$			
		BR_{MC}	V_{MC}	MSE_{MC}	CR	BR_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y, cal, \alpha}$	0.1393	4.8403	4.8500	0.956	0.1603	2.8293	2.8493	0.922
	$\hat{Q}_{y, HT, \alpha}$	-0.0477	5.8276	5.8182	0.934	-0.0227	3.5939	3.5872	0.924
	$\hat{Q}_{y, ra, \alpha}$	0.1648	9.5171	9.5252	0.980	0.1263	4.8687	4.8749	0.972
	$\hat{Q}_{y, diff, \alpha}$	-0.1418	4.7045	4.7152	0.960	-0.0464	2.9213	2.9176	0.936
	$\hat{Q}_{y, CD, \alpha}$	3.9150	3.5279	18.8477	0.584	3.9114	1.9163	17.2112	0.194
0.5	$\hat{Q}_{y, cal, \alpha}$	-0.1746	8.2437	8.2577	0.944	-0.2413	3.6477	3.6986	0.940
	$\hat{Q}_{y, HT, \alpha}$	-0.2824	10.1117	10.1712	0.916	-0.3343	4.5023	4.6050	0.936
	$\hat{Q}_{y, ra, \alpha}$	0.6558	50.4938	50.8228	0.944	0.4263	26.5883	26.7169	0.948
	$\hat{Q}_{y, diff, \alpha}$	-0.5975	17.0315	17.3544	0.972	-0.3496	8.9060	9.0104	0.970
	$\hat{Q}_{y, CD, \alpha}$	4.3173	4.4061	23.0363	0.484	4.0937	2.0711	18.8252	0.184
0.75	$\hat{Q}_{y, cal, \alpha}$	-0.2229	12.1861	12.2114	0.942	-0.2113	6.5823	6.6138	0.952
	$\hat{Q}_{y, HT, \alpha}$	-0.4150	14.2935	14.4371	0.934	-0.2786	7.6597	7.7220	0.934
	$\hat{Q}_{y, ra, \alpha}$	0.7861	47.3844	47.9077	0.980	-0.1344	19.5992	19.5781	0.958
	$\hat{Q}_{y, diff, \alpha}$	0.4347	52.3845	52.4687	0.972	-0.3409	23.8277	23.8962	0.958
	$\hat{Q}_{y, CD, \alpha}$	4.4114	7.7023	27.1478	0.654	4.3549	4.1566	23.1136	0.392

Table 8

Monte Carlo Simulation Results for Sampling from the SLID982 Population, Under PO Sampling Plan and the Second Rule for the Construction of the $\pi_k, k \in U$. The Number of Replications is Set at $K = 500$

α	Estimator	$n = 100$				$n = 200$			
		BR_{MC}	V_{MC}	MSE_{MC}	CR	BR_{MC}	V_{MC}	MSE_{MC}	CR
0.25	$\hat{Q}_{y, cal, \alpha}$	0.2392	3.4402	3.4906	0.962	0.1674	1.5214	1.5464	0.952
	$\hat{Q}_{y, HT, \alpha}$	0.0267	4.0027	3.9954	0.940	-0.0370	1.6995	1.6975	0.958
	$\hat{Q}_{y, ra, \alpha}$	0.4402	7.4350	7.6139	0.970	0.1850	3.0687	3.0968	0.978
	$\hat{Q}_{y, diff, \alpha}$	0.0528	3.2842	3.2804	0.972	-0.0127	1.4718	1.4690	0.964
	$\hat{Q}_{y, CD, \alpha}$	2.1458	3.0460	7.6444	0.876	1.9785	1.3010	5.2130	0.690
0.5	$\hat{Q}_{y, cal, \alpha}$	-0.1410	6.5627	6.5695	0.942	-0.2850	2.9662	3.0415	0.954
	$\hat{Q}_{y, HT, \alpha}$	-0.2133	7.6604	7.6906	0.928	-0.2876	3.6017	3.6772	0.926
	$\hat{Q}_{y, ra, \alpha}$	1.0245	43.2773	44.2402	0.930	-0.3075	17.7242	17.7833	0.948
	$\hat{Q}_{y, diff, \alpha}$	-0.1973	14.5261	14.5360	0.958	-0.6111	6.2988	6.6596	0.978
	$\hat{Q}_{y, CD, \alpha}$	2.2140	4.5617	9.4543	0.834	1.8882	2.0393	5.6005	0.738
0.75	$\hat{Q}_{y, cal, \alpha}$	-0.1985	12.6334	12.6476	0.952	-0.0022	5.6442	5.6329	0.966
	$\hat{Q}_{y, HT, \alpha}$	-0.4012	13.5045	13.6384	0.922	-0.1078	6.2239	6.2231	0.934
	$\hat{Q}_{y, ra, \alpha}$	0.7968	44.0650	44.6118	0.958	0.3727	19.1830	19.2836	0.960
	$\hat{Q}_{y, diff, \alpha}$	0.4613	49.6620	49.7755	0.960	0.2340	22.1292	22.1397	0.966
	$\hat{Q}_{y, CD, \alpha}$	2.6329	9.6723	16.5850	0.854	2.6729	4.1179	11.2541	0.738

5. Concluding Remarks

In this paper, we have developed quantile estimators based on the calibration paradigm. The estimators are particularly easy to implement and to interpret, since they focus on weights and calibration constraints. Furthermore, they require only the population quantiles of the auxiliary variables, which can be vectorial. When the quadratic metric is adopted, analytic expressions can be obtained for calibrated weights as well as variance estimators, which are similar to those for the calibration estimator for totals. From a practical point of view, an appealing consequence of the new methodology is that the proposed estimators are easy to calculate; it suffices to transform the auxiliary variables and then use existing software to compute the calibration estimators.

In a small simulation study, we compared the calibration estimator for quantiles, under the quadratic metric, to other leading quantile estimators available in the literature. The proposed estimator performed reasonably well in our empirical experiments; its performance was often preferable or at least similar to that of other estimators using the same amount of information. The model-based estimator incorporating much more information about the auxiliary variables appeared preferable under SRS sampling and a correctly specified model, but was outperformed by the new estimator when the first order inclusion probabilities were unequal. In general, the proposed estimator compared very well with the design-based alternatives of Rao *et al.* (1990).

While, in this paper, we have concentrated on the estimation of quantiles by calibrating on known population quantiles for the auxiliary variables, calibration estimators can be extended to other important estimation problems of interest in survey sampling. The formulation of these problems all lead to different transformed variables, that we have noted \mathbf{a}_k in this paper. For example, it is possible to formulate a calibration problem for the well-known Gini coefficient and then show that the solution to this calibration problem will give weights analogous to those derived in this paper; however these weights can only be determined numerically. More work is needed in this direction, in order to extend calibration estimators to a more general framework, which would include totals, quantiles, and Gini coefficients as special cases. Another challenging research avenue concerns the choice of the distribution function estimator. In this paper, we have advocated a distribution function estimator calculated using a linear interpolation. Alternatively, we could consider kernel distribution function estimator (see *e.g.*, Altman and Léger (1995)). Kernel density estimation from complex surveys is elaborated in Bellhouse and Stafford (1999). This means that, in $\hat{F}_{y, \text{cal}}(t)$, the function $H_{y, s}(t, y_k)$ could be replaced by a

general kernel, which would, however, depend on an additional parameter, the bandwidth. Note that the linear interpolation in the present paper avoids the choice of a bandwidth, which is often a delicate matter. Developing a general framework for calibration problems of a certain functional, and kernel distribution function estimators, are left for future studies.

Acknowledgements

The authors thank two anonymous referees for their thoughtful comments and suggestions, which greatly enhanced the paper. Discussions and comments from Raymond Chambers, Christian Léger, Éric Rancourt, Ulrich Rendtel and participants in the 32nd meeting of the Statistical Society of Canada and of the 2004 Joint Statistical Meeting are gratefully acknowledged. The first author was supported by a scholarship from the German Academic Exchange Service (DAAD) and the second author by grants from the National Science and Engineering Research Council of Canada and the Fonds québécois de la recherche sur la nature et les technologies du Québec (Canada).

References

- Altman, N., and Léger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46, 195-214.
- Bellhouse, D.R., and Stafford, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.
- Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika*, 63, 615-620.
- Chambers, R.L., Dorfman, A.H. and Hall, P. (1992). Properties of estimators of finite population distribution functions. *Biometrika*, 79, 577-582.
- Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.
- Chen, J., and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective use of auxiliary information. *Biometrika*, 80, 107-116.
- Déville, J.-C. (1988). Estimation linéaire et redressement sur information auxiliaire d'enquêtes par sondage. In *Essais en l'Honneur d'Edmont Malinvaud*, (Eds, A. Monfort, and J.J. Laffond), *Economica*, Paris, 915-929.
- Déville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Dorfman, A.H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics*, 35, 29-41.
- Harms, T. (2003). Extensions of the calibration approach: calibration of distribution functions and its link to small area estimators, Chintex working paper #13, Federal Statistical Office, Germany.

- Kovačević, M.S. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 139-144.
- Kovar, J.G., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16 (Supp.), 25-45.
- Kuk, A.Y.C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika*, 75, 97-103.
- Kuk, A.Y.C., and Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Series B (Methodological)*, 51, 261-269.
- Meeden, G. (1995). Median estimation using auxiliary information. *Survey Methodology*, 21, 71-77.
- Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.
- Ren, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. *Actes des journées de méthodologie statistique, INSEE Méthodes*, Tome 1, 100, 263-289.
- Ren, R., and Deville, J.C. (2000). Une généralisation du calage: calage sur les rangs et le calage sur les moments, II^{ème} Colloque Francophone sur les Sondages. Bruxelles.
- Rueda, M.M., Arcos A. and Martínez, M.D. (2003). Difference estimators of quantiles in finite populations. *Test*, 12, 481-496.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.
- Singh, A.C., and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.
- Thompson, M. (1997). *Theory of Sample Surveys*. Chapman & Hall, New York.
- Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 625-646.
- Wu, C., and Sitter, R.R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *The Canadian Journal of Statistics*, 29, 289-308.

A Nonresponse Model Approach to Inference Under Imputation for Missing Survey Data

David Haziza and Jon N.K. Rao¹

Abstract

In the presence of item nonresponse, two approaches have been traditionally used to make inference on parameters of interest. The first approach assumes uniform response within imputation cells whereas the second approach assumes ignorable response but make use of a model on the variable of interest as the basis for inference. In this paper, we propose a third approach that assumes a specified ignorable response mechanism without having to specify a model on the variable of interest. In this case, we show how to obtain imputed values which lead to estimators of a total that are approximately unbiased under the proposed approach as well as the second approach. Variance estimators of the imputed estimators that are approximately unbiased are also obtained using an approach of Fay (1991) in which the order of sampling and response is reversed. Finally, simulation studies are conducted to investigate the finite sample performance of the methods in terms of bias and mean square error.

Key Words: Bias-adjusted estimator; Deterministic regression imputation; Imputation model approach; Item nonresponse; Nonresponse model approach; Random regression imputation; Variance estimation.

1. Introduction

Item nonresponse occurs in a survey when a sampled element participates in the survey but fails to provide responses on one or more of the survey items (Brick and Kalton 1996). It is usually handled by some form of imputation which involves “filling in” missing values for each item. Imputation may achieve an effective bias reduction, provided suitable auxiliary information is available for all the sampled elements and appropriately incorporated in the imputation model and/or the non-response model.

Imputation offers the following desirable features, among others: (i) it leads to the creation of a complete data file, and (ii) it permits the use of the same survey weights for all items which ensures that the results obtained from different analyses of the completed data set are consistent with one another, unlike the results of analyses from an incomplete data set. However, imputation also presents the following difficulties, among others: (a) marginal imputation for each item distorts the relationship between items, and (b) treating the imputed values as if they were true values may lead to serious underestimation of the variance of imputed estimators, especially when the nonresponse rate is appreciable. Methods that address (a) and (b) have been proposed in the literature.

In this paper, we focus on marginal imputation that is commonly used in many surveys. We first consider deterministic linear regression imputation that includes mean and ratio imputation as special cases. In this method a missing value is replaced by the predicted value obtained by fitting a

linear regression model using respondent values and auxiliary variables collected on all the sampled elements. We also consider the case of random linear regression imputation that may be viewed as a deterministic regression imputation plus an added random residual. It includes random hot-deck imputation as a special case.

Let U be a finite population of possibly unknown size N . The objective is to estimate the population total $Y = \sum_U y_i$ of an item y when imputation has been used to compensate for nonresponse on the item values y_i . For brevity, \sum_A will be used for $\sum_{i \in A}$, where $A \subseteq U$. Suppose a probability sample, s , of size n is selected according to a specified design $p(s)$ from U . Under complete response to item y , a design-unbiased estimator of Y is given by the well-known Horvitz-Thompson estimator

$$\hat{Y}_I = \sum_s w_i y_i, \quad (1)$$

with sampling (or design) weights $w_i = 1/\pi_i$, where π_i denotes the inclusion probability of population unit i in the sample s , $i = 1, \dots, N$. Rao (2005) suggested that (1) should be called the Narain-Horvitz-Thompson (NHT) estimator in recognition of the fact that Narain (1951) also discovered (1) independently of Horvitz and Thompson (1952).

In the presence of nonresponse to item y , we use imputation and define an imputed estimator \hat{Y}_I^* as

$$\hat{Y}_I^* = \sum_s w_i a_i y_i + \sum_s w_i (1 - a_i) y_i^* = \sum_s w_i \tilde{y}_i, \quad (2)$$

where y_i^* denotes the value imputed for missing y_i , a_i denotes the response indicator equal to 1 if unit i responds to item y and 0 otherwise and $\tilde{y}_i = a_i y_i + (1 - a_i) y_i^*$. The

1. David Haziza, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada. K1S 5B6.

imputed estimator (2) can be implemented from the imputed data file containing the survey weights w_i and the \tilde{y}_i only, without the knowledge of response indicators a_i . However, the response indicators will be required for variance estimation. Let $p_i = P(a_i = 1)$ be the item y response probability for unit i . In this paper, we assume that the units respond independently of one another, i.e., $p_{ij} = P(a_i = 1, a_j = 1) = p_i p_j$ if $i \neq j$.

As for any method of compensating for missing data, imputation requires some assumptions about the response mechanism and/or the imputation model. In the presence of imputed data, two different approaches are generally used for making inference on totals, means and other parameters of interest: (i) Imputation model (IM) approach; (ii) Nonresponse Model (NM) approach. Approach (i) is also called model-assisted approach (Särndal 1992) and approach (ii) design-based approach (Shao and Steel 1999). NM approach is based on partitioning the population U into J imputation cells and then imputing nonrespondents y -values within each cell using respondent y -values within the same cell as donor values, independently across the J cells. The following assumption is made:

Assumption NM: Response probability for a given item of interest is constant within imputation cells. That is, $p_i = p_v$, say, where the subscript v denotes the imputation cell.

In the NM approach, explicit assumptions about the response mechanism are made. It follows that inference under assumption NM is with respect to repeated sampling and uniform response mechanism within cells. Approach NM has been studied by Rao (1990, 1996), Rao and Shao (1992), Rao and Sitter (1995) and Shao and Steel (1999), among others. For simplicity, we assume a single imputation cell so that $p_i = p$ under assumption NM.

IM approach is based on the following assumption:

Assumption IM: Item values are missing at random (MAR) in the sense that the response probability does not depend on the item value being imputed but may depend on auxiliary variables used for imputation. Further, a model that generates the item values y_i is assumed.

In the IM approach, explicit assumptions about the distribution of item values y_i is made through a model called the "imputation model". It follows that inference under assumption IM is with respect to repeated sampling and the assumed model that generates the finite population of y -values and nonrespondents to item y . Underlying response mechanism is not specified, except for the MAR assumption, unlike in the NM approach. The assumed response mechanism under assumption IM is much weaker than the uniform response within cells under assumption NM, but inferences under assumption IM depends on the

assumed population model. IM approach has been studied by Särndal (1992), Deville and Särndal (1994) and Shao and Steel (1999), among others.

Under linear regression imputation, IM approach assumes the following linear regression imputation model:

$$E_m(y_i) = \mathbf{z}_i' \boldsymbol{\gamma}, \quad V_m(y_i) = \sigma_i^2 = \sigma^2 (\boldsymbol{\lambda}' \mathbf{z}_i), \\ \text{Cov}_m(y_i, y_j) = 0 \text{ if } i \neq j, \quad (3)$$

where $\boldsymbol{\gamma}$ is k -vector of unknown parameters, \mathbf{z}_i is a k -vector of auxiliary variables available for all $i \in s$, $\boldsymbol{\lambda}$ is a k -vector of specified constants, σ^2 is an unknown parameter and E_m, V_m , and Cov_m denote respectively the expectation, the variance and the covariance operators with respect to the imputation model. The restriction $\sigma_i^2 = \sigma^2 (\boldsymbol{\lambda}' \mathbf{z}_i)$ does not severely restrict the range of imputation models.

In this paper, we propose a third approach, called the Generalized Nonresponse Model (GNM) approach. GNM approach is based on the following assumption:

Assumption GNM: Item values are missing at random (MAR) and response probability is specified as a function of auxiliary variables, \mathbf{u}_i , observed on all the sample elements, and unknown parameters $\boldsymbol{\eta}$.

In this paper, we assume that the probability of response, p_i , for unit i , is linked to an l -vector of auxiliary variables \mathbf{u}_i according to a logistic model so that

$$p_i = f(\mathbf{u}_i' \boldsymbol{\eta}) = \exp(\mathbf{u}_i' \boldsymbol{\eta}) / \exp(1 + \mathbf{u}_i' \boldsymbol{\eta}), \quad (4)$$

where $\boldsymbol{\eta}$ is the l -vector of model parameters. Model (4) is the assumed nonresponse model. It can be validated from the values a_i and \mathbf{u}_i for $i \in s$. Note that a_i and \mathbf{u}_i are item specific. Also, note that assumption NM is a special case of assumption GNM. As in NM approach, explicit assumptions about the response mechanism are made and inference under assumption GNM is with respect to repeated sampling and the assumed response mechanism.

Recall that imputation is designed to reduce the nonresponse bias, assuming that the available auxiliary variables can explain the item to be imputed and/or the item response probability. Hence, in practice, the choice of the approach (IM or GNM) should be dictated by the quality of the imputation model and the nonresponse model. The choice between modeling the item response probability and modeling the item of interest will depend on how much reliance one is ready to place on the two models. Although it may seem intuitively more appealing to model the item of interest, there are some cases encountered in practice for which it may be easier to model the response probability (GNM approach). For example, the Capital Expenditures Survey at Statistics Canada produces data on investment made in Canada, in all types of Canadian industries. For this survey, two important variables of interest are capital

expenditures on new construction (CC) and capital expenditures on new machinery and new equipment (CM). In a given year, a large number of businesses have not invested any amount of money on new construction or new machinery. As a result, the sample data file contains a large number of zeros for the two variables CC and CM. In this case, modeling the variables of interest (CC or CM) may prove to be difficult.

Survey design weights are generally used in linear regression imputation. The resulting imputed estimator of a population total is "robust" in the sense that it is approximately unbiased under either assumption NM or assumption IM. However, the imputed estimator is generally biased under assumption GNM. In this paper, we propose a new method of linear regression imputation that is robust in the sense of leading to approximately unbiased estimators under either assumption GNM or assumption IM.

Section 2 develops a new method of deterministic linear regression imputation as well as random linear regression imputation, and demonstrates the robustness property in estimating a population total Y . Results of a simulation study on the finite-sample performance of the imputed estimator under the new method of imputation are reported in section 3. Variance estimators are derived in section 4, using the 'reverse' approach of Fay (1991) in which the order of sampling and response is reversed:

Population \rightarrow census with nonrespondents \rightarrow sample
with nonrespondents.

Simulation results on variance estimators are also given. Finally, the case of domain means is investigated in section 5.

2. Estimation of a Total

In this section, we study the bias of the imputed estimator \hat{Y}_I . The total error, $\hat{Y}_I - Y$, may be decomposed as

$$\hat{Y}_I - Y = (\hat{Y} - Y) + (\hat{Y}_I - \hat{Y}). \quad (5)$$

The term $\hat{Y} - Y$ in (5) is called the sampling error, whereas the term $\hat{Y}_I - \hat{Y}$ is called the nonresponse/imputation error. Note that there is no imputation error under deterministic imputation. Since the sampling error does not depend on nonresponse and imputation method, we focus on the nonresponse/imputation error $\hat{Y}_I - \hat{Y}$ and evaluate its properties conditionally on the sample s . Under the NM or GNM approach, the conditional nonresponse bias is defined as $E_r(\hat{Y}_I - \hat{Y} | s)$, where $E_r(\cdot)$ denotes the expectation with respect to the response mechanism. Under the IM approach, the conditional nonresponse bias is defined as $E_r E_m(\hat{Y}_I - \hat{Y} | s)$ under MAR assumption.

2.1 Deterministic Regression Imputation

Deterministic regression imputation uses the imputed values

$$y_i^* = \mathbf{z}_i' \hat{\gamma}_r \quad (6)$$

for missing y_i , where

$$\hat{\gamma}_r = \left(\sum_s w_i a_i \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i) \right)^{-1} \sum_s w_i a_i \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i) \quad (7)$$

is the weighted least squares estimator of γ in the model (3), based on the sample elements responding to item y . Using (6), the imputed estimator (2) can be written as

$$\hat{Y}_I = \hat{Y}_r + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r)' \hat{\gamma}_r, \quad (8)$$

where $\hat{Y}_r = \sum_s w_i a_i y_i$, $\hat{\mathbf{Z}} = \sum_s w_i \mathbf{z}_i$ and $\hat{\mathbf{Z}}_r = \sum_s w_i a_i \mathbf{z}_i$. Note that the imputed estimator (8) is similar to a regression estimator in the case of two-phase sampling.

Under assumption NM, $E_r(a_i | s) = p$ and the conditional nonresponse bias, $E_r(\hat{Y}_I - \hat{Y} | s)$, is approximately equal to 0. Furthermore, under assumption IM and regression model (3), the conditional nonresponse bias $E_r E_m(\hat{Y}_I - \hat{Y} | s)$, is equal to 0. However, under assumption GNM, the conditional nonresponse bias is given by

$$E_r(\hat{Y}_I - \hat{Y} | s) \approx - \sum_s w_i (1 - p_i) (y_i - \mathbf{z}_i' \hat{\gamma}_p) \equiv B(\hat{Y}_I | s), \quad (9)$$

where

$$\hat{\gamma}_p = \left(\sum_s w_i p_i \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i) \right)^{-1} \sum_s w_i p_i \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i). \quad (10)$$

This result follows from the fact that under assumption GNM, $E_r(a_i | s) = p_i$. Hence, the choice of imputed values (6) is, in general, not suitable under assumption GNM. For the special case of assumption NM with $p_i = p$, the last term in (9) vanishes, noting that $(\sum_s w_i \mathbf{z}_i \mathbf{z}_i') \hat{\gamma}_p = \lambda' (\sum_s w_i \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i)) \hat{\gamma}_p = \lambda' (\sum_s w_i \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i)) = \sum_s w_i y_i$.

2.2 A Bias-Adjusted Estimator

We assume for now that the response probabilities p_i are known. A natural approach for eliminating the bias of \hat{Y}_I under assumption GNM is to consider a bias-adjusted estimator of the form

$$\hat{Y}_I^a = \hat{Y}_I - \hat{B}(\hat{Y}_I | s), \quad (11)$$

where $\hat{B}(\hat{Y}_I | s)$ is an estimator of $B(\hat{Y}_I | s)$:

$$\hat{B}(\hat{Y}_I | s) = - \sum_s w_i a_i \frac{(1 - p_i)}{p_i} (y_i - \mathbf{z}_i' \hat{\gamma}_r). \quad (12)$$

Note that $E_r[\hat{B}(\hat{Y}_I | s) | s] \approx B(\hat{Y}_I | s)$ under assumption GNM. Substituting (12) in (11), we get a bias-adjusted estimator as

$$\hat{Y}_I^a = \sum_s \frac{w_i}{p_i} a_i y_i + \left(\sum_s w_i \mathbf{z}_i' - \sum_s \frac{w_i}{p_i} a_i \mathbf{z}_i' \right) \hat{\gamma}_r. \quad (13)$$

Note that (13) is also in the form of a two phase regression estimator.

In practice, response probabilities p_i are unknown. Suppose we can obtain estimators \hat{p}_i of p_i by modelling p_i according to the nonresponse model (4). Then, a bias-adjusted estimator is obtained by replacing p_i in (13) with \hat{p}_i . This estimator is also approximately conditionally unbiased under assumption IM. Hence, the bias-adjusted estimator (13) is robust in the sense of validity under either assumption IM or assumption GNM. However, unlike the imputed estimator \hat{Y}_I given by (2), the bias-adjusted estimator \hat{Y}_I^a cannot be computed without the knowledge of the response identifiers, a_i , and the estimated response probabilities, \hat{p}_i . Hence, both the response indicators and the estimated response probabilities must be provided with the imputed data file to implement \hat{Y}_I^a , which may not be the case in practice. This drawback of \hat{Y}_I^a can be eliminated by using the new imputation method, given in section 2.3, that leads to an approximately unbiased estimator under either assumption GNM or assumption IM without the knowledge of a_i and \hat{p}_i on the imputed data file. However, for variance estimation, access to a_i and \hat{p}_i is needed.

2.3 Modified Deterministic Regression Imputation

We assume for now that the response probabilities p_i are known. We then use the imputed values

$$y_i^* = \mathbf{z}_i' \tilde{\gamma}_s \quad (14)$$

for missing y_i and obtain the form of $\tilde{\gamma}_s$ that leads to an approximately unbiased estimator under assumption GNM.

2.3.1 Approximately Unbiased Estimator

The following lemma gives the form of $\tilde{\gamma}_s$ that leads to an approximately unbiased estimator under assumption GNM.

Lemma 1: Under assumption GNM, the choice of $\tilde{\gamma}_s$ that leads to $E_r(\hat{Y}_I - \hat{Y} | s) = 0$ is given by

$$\tilde{\gamma}_{s,N} = \left[\sum_s w_i (1 - p_i) \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i) \right]^{-1} \sum_s w_i (1 - p_i) \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i). \quad (15)$$

Proof: The conditional nonresponse bias of \hat{Y}_I with $y_i^* = \mathbf{z}_i' \tilde{\gamma}_s$ under assumption GNM is given by

$$E_r(\hat{Y}_I - \hat{Y} | s) = - \sum_s w_i (1 - p_i) (y_i - \mathbf{z}_i' \tilde{\gamma}_s).$$

Noting that $(\lambda' \mathbf{z}_i) / (\lambda' \mathbf{z}_i) = 1$, it follows that $E_r(\hat{Y}_I - \hat{Y} | s) = 0$ if $\tilde{\gamma}_s$ satisfies

$$\lambda' \left[\sum_s w_i (1 - p_i) \mathbf{z}_i (y_i - \mathbf{z}_i' \tilde{\gamma}_s) / (\lambda' \mathbf{z}_i) \right] = 0. \quad (16)$$

The choice $\tilde{\gamma}_s = \tilde{\gamma}_{s,N}$ satisfies (16).

Note that $\tilde{\gamma}_{s,N}$ is unknown since the y -values are only observed for $i \in s_r$ and the response probabilities p_i are unknown. An estimator of $\tilde{\gamma}_{s,N}$, based on the responding units and estimated response probabilities \hat{p}_i , is given by

$$\tilde{\gamma}_r = \left[\sum_s w_i a_i \frac{(1 - \hat{p}_i)}{\hat{p}_i} \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i) \right]^{-1} \sum_s w_i a_i \frac{(1 - \hat{p}_i)}{\hat{p}_i} \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i). \quad (17)$$

We have $E_r(\tilde{\gamma}_r | s) \approx \tilde{\gamma}_{s,N}$ so that $\tilde{\gamma}_r$ is conditionally approximately unbiased for $\tilde{\gamma}_{s,N}$ under assumption GNM. Hence, using the imputed values

$$y_i^* = \mathbf{z}_i' \tilde{\gamma}_r \quad (18)$$

in (2) with $\tilde{\gamma}_r$ given by (17), leads to an approximately unbiased estimator of the total Y under assumption GNM. Note that $\tilde{\gamma}_r$ is a weighted least square estimator of γ with respect to a new set of weights, $\tilde{w}_i / (\lambda' \mathbf{z}_i)$, where $\tilde{w}_i = w_i ((1 - \hat{p}_i) / \hat{p}_i)$. Hence, the procedure increases the weights w_i for those units with $\hat{p}_i < 1/2$ and decreases the weights for those units with $\hat{p}_i > 1/2$. The imputed estimator can be implemented from the imputed data file containing the sampling weights w_i and the \tilde{y}_i only; response identifiers a_i and estimated response probabilities, \hat{p}_i , are not required. However, a_i and \hat{p}_i are needed for variance estimation. Note that the producer of the imputed data file uses the information on a_i and \mathbf{u}_i to fit the response model (4) and generate the imputed values y_i^* given by (18).

The use of imputed values (18) also leads to an approximately unbiased estimator of Y under assumption IM. First, under the regression model (3), noting that $E_m(y_i | s) = \mathbf{z}_i' \gamma$ and $E_m(\tilde{\gamma}_r | s) = \gamma$, we have $E_m(\hat{Y}_I - \hat{Y} | s) = 0$ and $E_r E_m(\hat{Y}_I - \hat{Y} | s) = 0$ without specifying the underlying MAR response mechanism. Hence, the use of imputed values (18) leads to a robust imputed estimator in the sense of validity under both approaches. Finally, it is interesting to note that the imputed values (18) can also be obtained using the method of calibration imputation (Beaumont 2005). Calibration imputation consists of finding final imputed values as close as possible to original imputed values according to some distance function, subject to the calibration constraint.

Two particular cases of modified regression imputation (18) are of interest: (i) modified ratio imputation with $\mathbf{z}_i = z_i$ and $\lambda' \mathbf{z}_i = z_i$; (ii) modified mean imputation with $\mathbf{z}_i = 1$ and $\lambda' \mathbf{z}_i = 1$. In case (i), the imputed values (18) reduce to

$$y_i^* = \sum_s \frac{\tilde{w}_i a_i y_i}{\sum_s \tilde{w}_i a_i z_i} z_i. \quad (19)$$

In case (ii), the imputed values (18) reduce to

$$y_i^* = \frac{\sum_s \tilde{w}_i a_i y_i}{\sum_s \tilde{w}_i a_i}. \quad (20)$$

Under uniform response $p_i = p$, the imputed values (19) and (20) reduce to $(\sum_s w_i a_i y_i / \sum_s w_i a_i z_i)$ and $\bar{y}_r = \sum_s w_i a_i y_i / \sum_s w_i a_i$ respectively, which are the usual values that survey practitioners use for ratio and mean imputation (Rao and Sitter 1995).

2.3.2 Optimal Choice of \tilde{y}_s

We now turn to the “optimal” choice of \tilde{y}_s by minimizing the conditional mean square error of the imputed estimator \hat{Y}_I with $y_i^* = \mathbf{z}_i' \tilde{\mathbf{y}}_s$. The conditional mean square error of the imputed estimator \hat{Y}_I is given by

$$\begin{aligned} \text{MSE}_r(\hat{Y}_I | s) &= V_r(\hat{Y}_I | s) + [\text{Bias}(\hat{Y}_I | s)]^2 \\ &= \sum_s w_i^2 p_i (1 - p_i) (y_i - \mathbf{z}_i' \tilde{\mathbf{y}}_s)^2 \\ &\quad + \left[\sum_s w_i (1 - p_i) (y_i - \mathbf{z}_i' \tilde{\mathbf{y}}_s) \right]^2, \end{aligned} \quad (21)$$

where $V_r(\cdot | s)$ denotes the conditional nonresponse variance with respect to the response mechanism, given the sample s . We search for $\tilde{\mathbf{y}}_s$ that minimizes $\text{MSE}_r(\hat{Y}_I | s)$.

The optimal choice, $\tilde{\mathbf{y}}_{\text{opt}}$, of $\tilde{\mathbf{y}}_s$ is complex, but in the special case of ratio imputation, $\tilde{\mathbf{y}}_{\text{opt}}$ reduces to

$$\tilde{\mathbf{y}}_{\text{opt}} = \frac{\sum_s w_i (1 - p_i) y_i \sum_s w_i (1 - p_i) z_i + \sum_s w_i^2 p_i (1 - p_i) y_i z_i}{\left[\sum_s w_i (1 - p_i) z_i \right]^2 + \sum_s w_i^2 p_i (1 - p_i) z_i^2}. \quad (22)$$

Assume that the sampling weights w_i satisfy $\max(n/Nw_i) = O(1)$ and that a positive constant C exists such that $C < p_i$. Then,

$$\begin{aligned} \tilde{\mathbf{y}}_{\text{opt}} &= \frac{\sum_s w_i (1 - p_i) y_i}{\sum_s w_i (1 - p_i) z_i} + O\left(\frac{1}{n}\right) \\ &= \tilde{\mathbf{y}}_{s,N} + O\left(\frac{1}{n}\right). \end{aligned}$$

Hence, for large sample sizes, the choice $\tilde{\mathbf{y}}_{s,N}$ is nearly optimal for ratio imputation. Similarly, $\tilde{\mathbf{y}}_{s,N}$ is nearly optimal for mean imputation which is a special case of ratio imputation.

2.4 Random Regression Imputation

Random imputation can be viewed as deterministic imputation plus a random noise. Let s_r and s_m denote the sets of sample respondents and nonrespondents respectively, and let $e_j = (y_j - \mathbf{z}_j' \hat{\mathbf{y}}_r) / (\lambda' \mathbf{z}_j)^{1/2}$ be the standardized residuals for the respondents $j \in s_r$ under deterministic

regression imputation. Further, $e_i^* = e_j$ with $P(e_i^* = e_j) = w_j / \sum_s w_i a_i$ independently for each $i \in s_m$. Then, random regression imputation uses the imputed values $y_i^* = \mathbf{z}_i' \hat{\mathbf{y}}_r + e_i^*$, $i \in s_m$, where $e_i^* = (\lambda' \mathbf{z}_i)^{1/2} (e_i^* - \bar{e}_r)$ with $\bar{e}_r = \sum_s w_j a_j e_j / \sum_s w_j a_j$. Let $E_*(\cdot)$ denote the expectation with respect to the random imputation process. We have $E_*(e_i^*) = 0$ and $E_*(\hat{Y}_I)$ equals (8). Hence, the imputed estimator \hat{Y}_I is approximately unbiased under either assumption NM or assumption IM. It may be noted that random regression imputation covers random (weighted) hot-deck imputation as a special case. To see this, consider the mean imputation model $E_m(y_i) = \gamma$, $V_m(y_i) = \sigma^2$ and $\text{Cov}_m(y_i, y_j) = 0$, $i \neq j$. We have $\hat{\mathbf{y}}_r = \sum_s w_i a_i y_i / \sum_s w_i a_i = \bar{y}_r$, the weighted mean of the respondent y -values, and $e_j = y_j - \bar{y}_r$. Therefore, $y_i^* = \bar{y}_r + e_i^* = y_j$ corresponds to the respondent value y_j drawn at random with probability $w_j / \sum_s w_i a_i$.

The imputed estimator based on random regression imputation is asymptotically biased under assumption GNM. To obtain an approximately unbiased estimator for Y , we propose modified random regression imputation. Let $\tilde{e}_j = (y_j - \mathbf{z}_j' \tilde{\mathbf{y}}_r) / (\lambda' \mathbf{z}_j)^{1/2}$ and $\tilde{e}_i^* = \tilde{e}_j$ with $P(\tilde{e}_i^* = \tilde{e}_j) = \tilde{w}_j / \sum_s \tilde{w}_i a_i$ independently for each $i \in s_m$, where $\tilde{\mathbf{y}}_r$ is given by (17) and $\tilde{w}_i = w_i (1 - \hat{p}_i) / \hat{p}_i$. Then, modified random regression imputation uses the imputed values $y_i^* = \mathbf{z}_i' \tilde{\mathbf{y}}_r + \tilde{e}_i^*$, where $\tilde{e}_i^* = (\lambda' \mathbf{z}_i)^{1/2} (\tilde{e}_i^* - \bar{\tilde{e}}_r)$ with $\bar{\tilde{e}}_r = \sum_s \tilde{w}_j a_j \tilde{e}_j / \sum_s \tilde{w}_j a_j$. We have $E_*(\tilde{e}_i^*) = 0$ and $E_*(\hat{Y}_I)$ equals the imputed estimator under modified deterministic regression imputation. Hence, the imputed estimator \hat{Y}_I is approximately unbiased under either assumption GNM or assumption IM. For the special case of mean imputation model, we have $\tilde{\mathbf{y}}_r = \sum_s \tilde{w}_i a_i y_i / \sum_s \tilde{w}_i a_i$ and $y_i^* = y_j$ corresponds to the respondent value y_j drawn at random with probability $\tilde{w}_j / \sum_s \tilde{w}_i a_i$.

3. Simulation Studies

We performed two simulation studies to investigate the finite sample performance of the proposed deterministic modified regression and modified random regression imputation methods in terms of relative bias and relative root mean square error. The first simulation study compares the performance of the traditional deterministic regression imputation and the proposed modified deterministic regression imputation when the imputation model and/or the non-response model are not correctly specified. The second simulation study compares the performance of the imputed estimator obtained by using imputation classes based on the estimated response probabilities and weighted mean imputation (traditional) with the imputed estimator obtained by using the proposed modified deterministic regression imputation method.

3.1 Simulation Study 1

We generated a finite population of size $N = 1,000$ containing 3 variables: a variable of interest y and two auxiliary variables z_1 and z_2 . To do so, we first generated z_1 and z_2 independently from an exponential distribution with mean 4 and 30 respectively. Then the y -values were generated according to the regression model

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \epsilon_i,$$

where the ϵ_i 's are generated from a normal distribution with mean 0 and variance σ^2 . The values of the parameters γ_0 , γ_1 and γ_2 were respectively set to 20, 2 and 0.1 and the variance σ^2 was chosen to lead to a model R^2 -value approximately equal to 0.75. The objective is to estimate the population total $Y = \sum_U y_i$.

We generated $R = 5,000$ simple random samples without replacement of size $n = 100$ from the finite population. In each sample, nonresponse to item y was generated according to the following response mechanisms:

Mechanism 1: Response probability p_{1i} for unit i is given by the logistic regression model

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i}.$$

Mechanism 2: Response probability p_{2i} for unit i is given by the logistic regression model

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i.$$

The values of λ_0 and λ_1 were chosen to give an overall response rate approximately equal to 70%. The response indicators a_{1i} and a_{2i} were generated independently from a Bernoulli distribution with parameters p_{1i} and p_{2i} , respectively. Note that in the case of the nonresponse mechanism 2, the response mechanism is nonignorable in the sense that the probability of response depends on the variable of interest y .

To compensate for the nonresponse to item y , we used the traditional deterministic regression imputation for which the imputed values are given by (6) and the modified deterministic regression imputation for which the imputed values are given by (18). Imputations were based on the models for y and for p listed in Table 1 as $y_{(1)}$, $y_{(2)}$, $y_{(3)}$, $y_{(4)}$ and $p_{(1)}$, $p_{(2)}$, $p_{(3)}$. Note that $p_{(1)}$ corresponds to response mechanism 1 and $y_{(1)}$ to the model generating the population.

From each simulated sample, we calculated the imputed estimator \hat{Y}_I given by (2) with the imputed values (6) and (18), based on selected combinations of the models $y_{(a)}$ and $p_{(b)}$; $a = 1, \dots, 4$; $b = 1, 2, 3$. As a measure of the bias of an imputed estimator \hat{Y}_I , we used the percent simulated relative bias (RB) given by

$$RB(\hat{Y}_I) = \frac{\text{Bias}(\hat{Y}_I)}{Y} \times 100, \quad (23)$$

where

$$\text{Bias}(\hat{Y}_I) = \frac{1}{R} \sum_{r=1}^R \hat{Y}_I^{(r)} - Y \quad (24)$$

and $\hat{Y}_I^{(r)}$ denotes the value of \hat{Y}_I for the r -th simulated sample. As a measure of variability of an imputed estimator \hat{Y}_I , we used the percent simulated relative root mean square error (RRMSE) given by

$$\text{RRMSE}(\hat{Y}_I) = \frac{\sqrt{\text{MSE}(\hat{Y}_I)}}{Y} \times 100, \quad (25)$$

where

$$\text{MSE}(\hat{Y}_I) = \frac{1}{R} \sum_{r=1}^R (\hat{Y}_I^{(r)} - Y)^2. \quad (26)$$

Table 1
Models Used for Imputation

Models for y	Intercept	z_1	z_2
$y_{(1)}$	Yes	Yes	Yes
$y_{(2)}$	Yes	No	Yes
$y_{(3)}$	Yes	Yes	No
$y_{(4)}$	No	Yes	Yes
Models for p_i	Intercept	z_1	z_2
$p_{(1)}$	Yes	Yes	No
$p_{(2)}$	Yes	No	Yes
$p_{(3)}$	No	Yes	No

Results on relative bias and RRMSE are shown in Table 2 for the samples generated by response mechanism 1 and in Table 3 for the samples generated by the response mechanism 2. From Table 2, it is clear that, when the imputation is performed according to the correct model (*i.e.*, $y_{(1)}$), traditional deterministic regression imputation leads to an approximately unbiased estimator and it is more efficient than the modified deterministic regression imputation in terms of RRMSE. As noted by a referee, modified deterministic regression imputation can lead to more efficient estimators than traditional deterministic regression. That is, there are scenarios (not considered here) for which the proposed modified deterministic regression imputation method may be more efficient than the traditional deterministic regression imputation method.

When the imputation model is incorrectly specified (*e.g.*, $y_{(2)}$ and $y_{(4)}$), deterministic imputation leads to biased estimators whereas the bias of the modified deterministic imputation is small to negligible, provided the nonresponse model is correctly specified (*i.e.*, $p_{(1)}$). As a result, RRMSE for the deterministic imputation is larger than that for the

modified deterministic regression imputation. When both imputation and nonresponse models are not correctly specified (e.g., $y_{(4)} - p_{(2)}$), all the estimators are biased.

From Table 3, it is clear that, for the case of mechanism 2, the imputed estimator obtained under modified regression imputation performs equally or better than the imputed estimator obtained under traditional regression imputation in all the scenarios. This result is not suprising since achieving an effective bias reduction in the case of nonignorable nonresponse requires the use of all the appropriate auxiliary information available. The auxiliary information used in the case of the proposed modified regression imputation is richer than the one used in the case of regression imputation since it uses the auxiliary variables that are related to both the variable of interest y and the response probability whereas regression imputation uses only the auxiliary variables related to the variable of interest y .

Table 2
Relative Bias (%) and RRMSE (%) of Imputed Estimators Under Response Mechanism 1

Scenario	Bias (traditional)	Bias (proposed)	RRMSE (traditional)	RRMSE (proposed)
$y_{(1)} - p_{(1)}$	0.19	-0.01	1.85	2.33
$y_{(2)} - p_{(1)}$	5.20	0.16	5.60	2.66
$y_{(3)} - p_{(1)}$	0.17	-0.04	1.87	2.37
$y_{(4)} - p_{(1)}$	-14.80	-3.50	15.00	6.70
$y_{(1)} - p_{(2)}$	0.19	0.12	1.85	1.86
$y_{(4)} - p_{(2)}$	-14.80	-14.80	15.00	14.60
$y_{(1)} - p_{(3)}$	0.19	0.05	1.85	1.88

Table 3
Relative Bias (%) and RRMSE (%) of Imputed Estimators Under Response Mechanism 2

Scenario	Bias (traditional)	Bias (proposed)	RRMSE (traditional)	RRMSE (proposed)
$y_{(1)} - p_{(1)}$	1.84	1.83	2.55	2.54
$y_{(2)} - p_{(1)}$	4.46	1.84	4.89	2.65
$y_{(3)} - p_{(1)}$	2.03	2.02	2.70	2.70
$y_{(4)} - p_{(1)}$	-4.58	-3.04	5.07	3.81
$y_{(1)} - p_{(2)}$	1.84	1.84	2.55	2.55
$y_{(4)} - p_{(2)}$	-4.58	-1.70	5.07	2.88
$y_{(1)} - p_{(3)}$	1.84	1.84	2.55	2.55

3.2 Simulation Study 2

We generated a finite population of size $N = 1,000$ containing 3 variables: a variable of interest y and three auxiliary variables z_1, z_2 and z_3 , by first generating z_1, z_2 and z_3 independently from an exponential distribution with mean 100 and then generating the y -values according to the regression model

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i}^2 + \epsilon_i,$$

where the ϵ_i 's are generated from a normal distribution with mean 0 and variance σ^2 . The values of the parameters $\gamma_0, \gamma_1, \gamma_2$ and γ_3 were respectively fixed to 20, 10, 0.5 and 10. The variance σ^2 was chosen to lead to a model R^2 approximately equal to 0.66. The objective is to estimate the population mean $\bar{Y} = \sum_U y_i / N$. In order to focus on the nonresponse/imputation error, we considered the case of a census, i.e., $n = N = 1,000$. From the simulated population, nonresponse to item y was generated according to the following response mechanisms:

Mechanism 1: Response probability p_{1i} for unit i is given by the logistic model

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i} + \lambda_2 z_{3i}.$$

Mechanism 2: Response probability p_{2i} for unit i is given by the logistic model

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i + \lambda_2 z_{3i}.$$

The values of λ_0, λ_1 and λ_2 were chosen to give an overall response rate approximately equal to 70%. Response indicators a_{1i} and a_{2i} were then genrated independently $R = 1,000$ times from a Bernoulli distribution with parameters p_{1i} and p_{2i} , respectively.

To compensate for nonresponse, two strategies were used: The first strategy consisted in dividing the sample, s , into imputation classes s_1, s_2, \dots, s_C based on the auxiliary variables z_1, z_2 and z_3 . To form the classes, we used the score method which may be described as follows: Using the auxiliary information, we first estimated the response probabilites, p_i , to obtain \hat{p}_i for both the respondents and the nonrespondents using logistic regression on z_1, z_2 and z_3 . Using the \hat{p}_i 's, we then partitioned the population into C classes using the procedure FASTCLUS of SAS (that uses the k -means classification algorithm). The score method leads to a partition of the population in such a way that, within classes, units (respondents and nonrespondents) are homogeneous with respect to \hat{p}_i -values. The second strategy used the proposed modified regression imputation method based on the auxiliary variables z_1, z_2 and z_3 . The goal of the simulation study is to compare the performances of two imputed estimators of the population mean \bar{Y} : (a) Imputed estimator based on the C imputation classes:

$$\bar{y}_I^C = \sum_{c=1}^C \frac{\hat{N}_c}{\hat{N}} \bar{y}_{Ic}, \tag{27}$$

where

$$\bar{y}_{Ic} = \frac{1}{\hat{N}_c} \left[\sum_{s_c} w_i a_i y_i + \sum_{s_c} w_i (1 - a_i) y_i^* \right],$$

and $\hat{N}_c = \sum_{s_c} w_i$. We used weighted mean imputation within classes; i.e., $y_i^* = \sum_{s_c} w_i a_i y_i / \sum_{s_c} w_i a_i$.

(b) Imputed estimator based on the proposed modified regression imputation, denoted \bar{y}_I :

$$\bar{y}_I = \frac{1}{\hat{N}} \left[\sum_s w_i a_i y_i + \sum_s w_i (1 - a_i) y_i^* \right], \quad (28)$$

where the imputed values y_i^* are given by (18) using $\mathbf{z}_i' = (z_{1i}, z_{2i})'$ and $\hat{N} = \sum_s w_i$. For mechanism 1, the response probabilities p_i were correctly estimated using the variable z_1 and z_3 whereas the variables z_1, z_2 and z_3 were used to estimate p_i for mechanism 2.

Note that $w_i = 1$ in this simulation study for all $i \in U$ because no sampling is involved. Finally, Table 4 compares these estimators in terms of relative bias, given by (23) and RRMSE, given by (25). From Table 4, it is clear that the proposed imputed estimator (28) performs considerably better than the estimator (27) based on imputation classes in terms of RRMSE for both mechanism 1 and mechanism 2.

Table 4

Relative Bias (%) and RRMSE (%) of Imputed Estimators

Imputed estimator*	Number of classes	RB	RRMSE
\bar{y}_I^C (mechanism 1)	1	14.4	14.5
	5	-0.02	4.26
	10	-0.85	7.33
	20	-0.20	8.61
	30	-0.03	8.61
	40	0.03	9.09
\bar{y}_I (mechanism 1)	50	0.06	9.44
	—	1.11	1.90
\bar{y}_I^C (mechanism 2)	1	29.0	29.1
	5	21.4	21.4
	10	21.0	21.1
	20	20.9	21.0
	30	20.9	21.0
	40	21.0	21.0
\bar{y}_I (mechanism 2)	50	21.0	21.0
	—	10.9	10.9

* \bar{y}_I^C given by (27) and \bar{y}_I given by (28).

4. Variance Estimation

In this section, we derive a variance estimator of the imputed estimator \hat{Y}_I , using the reverse approach of Fay (1991). The total variance of \hat{Y}_I under a particular deterministic imputation method, is given by

$$V(\hat{Y}_I - Y) = E_r V_p(\hat{Y}_I - Y | \mathbf{a}) + V_r E_p(\hat{Y}_I - Y | \mathbf{a}), \quad (29)$$

where $\mathbf{a} = (a_1, \dots, a_N)'$ is the vector of response indicators, (Shao and Steel 1999). An estimator of the overall variance $V(\hat{Y}_I - Y)$ in (29) is given by $v_i = v_1 + v_2$, where v_1 is an estimator of $V_p(\hat{Y}_I - Y | \mathbf{a})$ conditional on the response indicators a_i , and v_2 is an estimator of $V_r[E_p(\hat{Y}_I - Y | \mathbf{a})]$. The estimator v_1 does not depend on the response

mechanism or the imputation model, and hence v_1 is valid under either assumption GNM or assumption IM.

Under the corresponding random imputation, the variance of the imputed estimator \hat{Y}_I is given by

$$V(\hat{Y}_I - Y) = E_r V_p E_s(\hat{Y}_I - Y | \mathbf{a}) + E_r E_p V_s(\hat{Y}_I - Y | \mathbf{a}) + V_r E_p E_s(\hat{Y}_I - Y | \mathbf{a}), \quad (30)$$

where $V_s(\cdot)$ denotes the variance operator with respect to random imputation. We assume that $E_s(\hat{Y}_I | \mathbf{a})$ agrees with the imputed estimator for the deterministic case. Hence, $E_r V_p E_s(\hat{Y}_I - Y | \mathbf{a})$ is estimated by v_1 for the deterministic case. Similarly, $V_r E_p E_s(\hat{Y}_I - Y | \mathbf{a})$ is estimated by v_2 for the deterministic case. The additional contribution to variance due to random imputation comes from the component $E_r E_p V_s(\hat{Y}_I - Y | \mathbf{a})$, which is estimated by $v_s = V_s(\hat{Y}_I - Y | \mathbf{a})$. Hence, it follows from (30) that the overall variance $V(\hat{Y}_I - Y)$ is estimated by $v_i = v_1 + v_s + v_2$. The term v_s is absent for deterministic imputation.

4.1 Known p_i

In this section, we assume that the response probabilities p_i are known. We first consider the case of modified deterministic regression imputation in section 4.1.1. The case of modified random regression imputation is studied in section 4.1.2.

4.1.1 Modified Deterministic Regression Imputation

Under modified deterministic regression imputation, the imputed estimator with known p_i may be written as

$$\hat{Y}_{lp} = \sum_s w_i a_i y_i + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r)' \tilde{\gamma}_{rp}, \quad (31)$$

where

$$\tilde{\gamma}_{rp} = \left[\sum_s w_i a_i \frac{(1 - p_i)}{p_i} \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i) \right]^{-1} \left[\sum_s w_i a_i \frac{(1 - p_i)}{p_i} \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i) \right]. \quad (32)$$

To obtain v_1 , we use standard Taylor linearization which leads to

$$\hat{Y}_{lp} - Y \approx \sum_s w_i \tilde{\xi}_{ip}, \quad (33)$$

where

$$\tilde{\xi}_{ip} = a_i y_i + (1 - a_i) \mathbf{z}_i' \tilde{\gamma}_{rp} + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r)' \tilde{\mathbf{T}}_p^{-1} a_i \frac{(1 - p_i)}{p_i} \frac{1}{(\lambda' \mathbf{z}_i)} \mathbf{z}_i (y_i - \mathbf{z}_i' \tilde{\gamma}_{rp})$$

with $\tilde{\mathbf{T}}_p = \sum_s w_i a_i ((1 - p_i) / p_i) \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i)$. Denoting the variance estimator of the full sample estimator as

$\hat{Y} = \sum_s w_i y_i$ as $v(y)$, it follows from (33) that an estimator of $V_p(\hat{Y}_I - Y | \mathbf{a})$ is given by

$$v_1 = v(\tilde{\xi}_p), \quad (34)$$

which is obtained by replacing y_i by \tilde{y}_{ip} in the formula for $v(y)$.

To obtain the second component v_2 , first note that

$$E_p(\hat{Y}_{lp} - Y | \mathbf{a}) \approx \sum_s a_i y_i + \sum_U (1 - a_i) \gamma_p - Y,$$

where

$$\gamma_p =$$

$$\left[\sum_U a_i \frac{(1 - p_i)}{p_i} \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i) \right]^{-1} \sum_U a_i \frac{(1 - p_i)}{p_i} \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i).$$

Using Taylor linearization, it can be shown that

$$V_r[E_p(\hat{Y}_{lp} - Y | \mathbf{a})] \approx \sum_U p_i (1 - p_i) \zeta_i^2, \quad (35)$$

where

$$\zeta_i = \left[1 + \frac{(1 - p_i)}{p_i} \frac{1}{(\lambda' \mathbf{z}_i)} (\mathbf{Z} - \mathbf{Z}_r)' \mathbf{T}_p^{-1} \mathbf{z}_i \right] (y_i - \mathbf{z}_i' \gamma_p)$$

with $\mathbf{Z} = \sum_U \mathbf{z}_i$, $\mathbf{Z}_r = \sum_U a_i \mathbf{z}_i$ and $\mathbf{T}_p = \sum_U a_i ((1 - p_i) / p_i) \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i)$. The component v_2 is then obtained by estimating the unknown quantities in (35), which leads to

$$v_2 = \sum_s w_i a_i (1 - p_i) \hat{\zeta}_i^2, \quad (36)$$

where

$$\hat{\zeta}_i = \left[1 + \frac{(1 - p_i)}{p_i} \frac{1}{(\lambda' \mathbf{z}_i)} (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r)' \hat{\mathbf{T}}_p^{-1} \mathbf{z}_i \right] (y_i - \mathbf{z}_i' \tilde{\gamma}_p).$$

An estimator of the total variance v_t is obtained as the sum of (34) and (36): $v_t = v_1 + v_2$. In practice, the response probabilities are unknown. As a result, it is not possible to calculate the variance estimator v_t . A simple solution consists in replacing p_i by the estimated response probabilities \hat{p}_i in (34) and (36) and use the resulting v_t as the variance estimator of \hat{Y}_I . As we show in a simulation study in section 4.3, this simple method gives acceptable results.

4.1.2 Modified Random Regression Imputation

We first note that

$$V_*(y_i^*) =$$

$$(\lambda' \mathbf{z}_i) \sum_s w_j \frac{(1 - p_j)}{p_j} a_i (\tilde{e}_j - \tilde{e}_r)^2 \left/ \sum_s w_j \frac{(1 - p_j)}{p_j} a_j \right. \equiv \tilde{s}_e^2$$

and $\text{Cov}_*(y_i^*, y_j^*) = 0, i \neq j$. Hence, from (2) the component v_* , due to random imputation, is given by

$$v_* = \sum_s w_i^2 (1 - a_i) V_*(y_i^*) = \sum_s w_i^2 (1 - a_i) \tilde{s}_e^2. \quad (37)$$

An estimator of the total variance is obtained as the sum of (34), (36) and (37): $v_t = v_1 + v_2 + v_*$. Once again, since the response probabilities p_i are unknown, it is not possible to

compute v_* in (37). We propose to replace p_i in (37) by the estimated response probabilities \hat{p}_i .

4.2 Unknown p_i

We use Binder's method (Binder 1983) to derive the component v_1 when the response probabilities p_i are estimated. We assume that $p_i = f(\mathbf{u}_i' \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is 1-vector of unknown parameters, \mathbf{u}_i is a 1-vector of auxiliary variables available for all $i \in s$. For example, in the case of logistic regression, $f(\mathbf{u}_i' \boldsymbol{\eta}) = \exp(\mathbf{u}_i' \boldsymbol{\eta}) / \exp(1 + \mathbf{u}_i' \boldsymbol{\eta})$. The estimated response probabilities are given by $\hat{p}_i = f(\mathbf{u}_i' \hat{\boldsymbol{\eta}})$, where $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$. Let $\boldsymbol{\theta} = (\boldsymbol{\eta}'_N, \boldsymbol{\gamma}'_N, Y')'$, where $\boldsymbol{\eta}_N$ and $\boldsymbol{\gamma}_N$ are census parameter corresponding to $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$, respectively. An estimator of $\boldsymbol{\theta}$ given by $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\eta}}', \hat{\boldsymbol{\gamma}}', \hat{Y}_I)'$ can be expressed as a solution of the sample estimating equations

$$\hat{\mathbf{S}}(\boldsymbol{\theta}) = \mathbf{0},$$

where $\hat{\mathbf{S}}(\boldsymbol{\theta}) = (\hat{\mathbf{S}}_1(\boldsymbol{\theta}), \hat{\mathbf{S}}_2(\boldsymbol{\theta}), \hat{\mathbf{S}}_3(\boldsymbol{\theta}))'$ with

$$\hat{\mathbf{S}}_1(\boldsymbol{\theta}) = \sum_s w_i \mathbf{u}_i [a_i - f(\mathbf{u}_i' \boldsymbol{\eta}_N)] = \mathbf{0},$$

$$\hat{\mathbf{S}}_2(\boldsymbol{\theta}) = \sum_s w_i a_i \mathbf{z}_i \frac{(1 - f(\mathbf{u}_i' \boldsymbol{\eta}_N))}{f(\mathbf{u}_i' \boldsymbol{\eta}_N)} (y_i - \mathbf{z}_i' \boldsymbol{\gamma}_N) / (\lambda' \mathbf{z}_i) = \mathbf{0}$$

and

$$\hat{\mathbf{S}}_3(\boldsymbol{\theta}) = Y - \sum_s w_i \mathbf{z}_i' \boldsymbol{\gamma}_N - \sum_s w_i a_i (y_i - \mathbf{z}_i' \boldsymbol{\gamma}_N) = 0.$$

Let $\hat{\mathbf{J}}(\boldsymbol{\theta}) = (\partial \hat{\mathbf{S}}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta})$ be the $(k + l + 1) \times (k + l + 1)$ matrix of partial derivative. We have

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})] \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}) [\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})]',$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ denotes the $(k + l + 1) \times (k + l + 1)$ symmetric matrix whose ij element is the covariance between $\hat{S}_i(\boldsymbol{\theta})$ and $\hat{S}_j(\boldsymbol{\theta})$ with respect to sampling given the vector of response indicator \mathbf{a} . If $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is replaced by a consistent estimator $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})$, say, we obtain a consistent variance estimator $\mathbf{v}(\hat{\boldsymbol{\theta}})$ given by

$$\mathbf{v}(\hat{\boldsymbol{\theta}}) = [\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})] \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}) [\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})]'.$$

Since we are interested in the variance estimator, v_1 , of \hat{Y}_I , we need the final row, \mathbf{b} , say, of $\hat{\mathbf{J}}^{-1}(\hat{\boldsymbol{\theta}})$, evaluated at $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$. It follows that

$$v_1 = \mathbf{b}' \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}}) \mathbf{b}'. \quad (38)$$

To obtain the component v_2 , we assume that the sampling weights w_i satisfy $\max(n / N w_i) = O(1)$ and that there exists a positive constant C such that $C < p_i$. Furthermore, we assume that $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta} = O_p(n^{-1/2})$. By Taylor linearization, we have

$$\hat{Y}_I = \hat{Y}_{lp} + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \sum_s p_i^{-1} (y_i - \tilde{y}_a) \frac{\partial f(\mathbf{u}_i' \boldsymbol{\eta})}{\partial \boldsymbol{\eta}} + O_p(N / n),$$

where

$$\tilde{\gamma}_a = \left[\sum_U (1 - a_i) \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i) \right]^{-1} \left[\sum_U (1 - a_i) \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i) \right].$$

Assuming that $f(\mathbf{u}'_i \boldsymbol{\eta}) / \partial \boldsymbol{\eta}$ is uniformly bounded, we have

$$E_p(\hat{Y}_I) = E_p(\hat{Y}_{Ip}) + O_p(N/n^{1/2}).$$

Hence, the component $V_r[E_p(\hat{Y}_{Ip} - Y | \mathbf{a})]$ is approximately given by (35) and v_2 is given by (36) with p_i replaced by \hat{p}_i . In the case of modified random regression imputation, the component due to random imputation will be estimated by (37) with p_i replaced by \hat{p}_i .

4.3 Simulation Study

We performed a limited simulation study to assess the performance of the variance estimators considered in sections 4.1 and 4.2. We generated a population of size $N = 2,500$ containing two variables y and z . First, the variable z was generated from a Gamma distribution with scale parameter equal to 4 and shape parameter equal to 10. The y -values were then generated according to the ratio model

$$y_i = \gamma z_i + \epsilon_i,$$

where the ϵ_i 's are generated from a normal distribution with mean 0 and variance σ^2 . The value of the parameter γ was set to 2 and the variance σ^2 was chosen to lead to a model R^2 -value approximately equal to 0.81. The objective is to estimate the population total $Y = \sum_U y_i$.

We generated $R = 10,000$ simple random samples without replacement from the finite population using the following sampling fractions n/N : 0.05; 0.1 and 0.25. In each sample, nonresponse to item y was generated according to the following response mechanism: Response probability p_i for unit i is given by the logistic model

$$\log \frac{p_i}{1 - p_i} = \lambda_0 + \lambda_1 z_i.$$

The values of λ_0 and λ_1 were chosen to give an overall response rate approximately equal to 70%. The response indicators a_i were then generated independently from a Bernoulli distribution with parameters p_i .

To compensate for the nonresponse to item y , we used the modified deterministic ratio imputation for which the imputed values are given by (19). From each simulated sample, we calculated the imputed estimator \hat{Y}_I given by (2) with the imputed values (19). As a measure of the bias of a variance estimator v , we used the relative bias $[E(v) - \text{MSE}(\hat{Y}_I)] / \text{MSE}(\hat{Y}_I)$. Let v_{naive} denotes the total variance estimator obtained by summing (34) and (36) when the response probabilities p_i are replaced by the estimated response probabilities \hat{p}_i and v_{correct} denotes the total variance estimator obtained by summing (38) and (36) with p_i replaced by \hat{p}_i . Table 5 gives the relative bias (in %) of

the two variance estimators. It is clear from Table 5 that both variance estimators lead to underestimation, but v_{correct} is slightly better in terms of underestimation. Also, both variance estimators performed well with a relative bias less than -10%. Hence, the simpler variance estimator v_{naive} might be suitable in practice.

Table 5
Relative Bias (%) of the Variance Estimators

f	RB(v_{naive})	RB(v_{correct})
0.05	-6.3	-5.1
0.10	-5.8	-4.1
0.25	-4.3	-3.2

5. Estimation of Domain Means

In practice, estimates for various domains (subpopulations) are often needed. For example, in the Canadian Labour Force Survey, estimates of unemployment are required by age-sex group and by industry at the provincial level. To compensate for item nonresponse, the proposed modified regression imputation may be used. However, the domains must be specified in advance at the imputation stage. In other words, the domain indicators must be part of the imputation model. In practice, domains are generally not specified at the edit and imputation stage and domain estimates are obtained from imputed data based on imputation models without the domain indicators. As a result, the imputed estimators for domains are generally biased. We propose a bias-adjusted estimator, along the lines of section 2.2, to remedy this problem. The bias-adjusted estimator can be obtained at the estimation stage and does not require the specification of the domains at the imputation stage.

A vector of domain means may be expressed as

$$\bar{\mathbf{Y}}_{(d)} = \left(\sum_U \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_U \mathbf{x}_i y_i, \quad (39)$$

where $\mathbf{x} = (x_{i1}, \dots, x_{di}, \dots, x_{Di})'$ is a vector of domain indicators, x_{di} , such that $x_{di} = 1$ if $i \in \text{domain } d$ and $x_{di} = 0$, otherwise. We assume that \mathbf{x} is known for all the units $i \in s$. In other words, only item y may be missing. In the absence of nonresponse, an approximately unbiased estimator of $\bar{\mathbf{Y}}_{(d)}$ is given by

$$\hat{\bar{\mathbf{Y}}}_{(d)} = \left(\sum_s w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_s w_i \mathbf{x}_i y_i. \quad (40)$$

In the presence of nonresponse to item y , an imputed estimator of $\bar{\mathbf{Y}}_{(d)}$ is given by

$$\begin{aligned} \hat{\bar{\mathbf{Y}}}_{I(d)} &= \hat{\mathbf{T}}^{-1} \left[\sum_s w_i a_i \mathbf{x}_i y_i + \sum_s w_i (1 - a_i) \mathbf{x}_i y_i^* \right] \\ &= \hat{\mathbf{T}}^{-1} \sum_s w_i a_i \mathbf{x}_i \tilde{y}_i, \end{aligned} \quad (41)$$

where $\hat{\mathbf{T}} = \sum_s w_i \mathbf{x}_i \mathbf{x}_i'$. Note that the imputed estimator $\hat{\mathbf{Y}}_{I(d)}$ in (41) does not require the response identifiers, a_i . Haziza and Rao (2005) showed that the imputed estimator $\hat{\mathbf{Y}}_{I(d)}$ is biased under assumption NM. They proposed a bias-adjusted estimator which is approximately unbiased under either assumption NM or assumption IM. In this section, we propose an extension of the Haziza-Rao bias-adjusted estimator which is approximately unbiased under either assumption GNM or assumption IM.

It is easily seen that, under assumption GNM, the conditional nonresponse bias of the imputed estimator (41) that uses the modified deterministic regression imputation (18) is given by

$$\text{Bias}(\hat{\mathbf{Y}}_{I(d)} | s) \approx -\hat{\mathbf{T}}^{-1} \left[\sum_s w_i (1 - p_i) \mathbf{x}_i (y_i - \mathbf{z}_i' \tilde{\gamma}_{s,N}) \right], \quad (42)$$

where $\tilde{\gamma}_{s,N}$ is given by (15). An approximately conditionally unbiased estimator of the bias in (42) is given by

$$\hat{B}(\hat{\mathbf{Y}}_{I(d)} | s) \approx -\hat{\mathbf{T}}^{-1} \left[\sum_s \tilde{w}_i a_i \mathbf{x}_i (y_i - \mathbf{z}_i' \tilde{\gamma}_r) \right], \quad (43)$$

where $\tilde{\gamma}_r$ is given by (17). A bias-adjusted estimator, $\hat{\mathbf{Y}}_{I(d)}^a$, is then obtained as $\hat{\mathbf{Y}}_{I(d)} - \hat{B}(\hat{\mathbf{Y}}_{I(d)} | s)$, which leads to

$$\hat{\mathbf{Y}}_{I(d)}^a = \hat{\mathbf{T}}^{-1} \left[\sum_s \frac{w_i}{\hat{p}_i} a_i \mathbf{x}_i (y_i - \mathbf{z}_i' \tilde{\gamma}_r) + \sum_s w_i \mathbf{x}_i \mathbf{z}_i' \tilde{\gamma}_r \right]. \quad (44)$$

The bias-adjusted estimator (44) is approximately unbiased under either IM or GNM. Hence, it is robust in the sense of validity under both assumption IM or assumption GNM. However, it requires both the response identifiers a_i and the estimated response probabilities \hat{p}_i , unlike the imputed estimator $\hat{\mathbf{Y}}_{I(d)}$ in (41).

It is possible to obtain a bias-adjusted estimator of the form (44) if we use the traditional deterministic regression imputation instead. It is interesting to note that the bias-adjusted estimator is identical to the estimator obtained using calibrated imputation (Beaumont 2005). The latter estimator does not require the knowledge of a_i and \hat{p}_i in the imputed data file but the domains must be specified at the imputation stage, which may not be feasible in practice.

If the nonresponse model (4) contains only the intercept, we have $\hat{p}_i = \hat{p}$, where \hat{p} denotes the overall response rate. In this case, the bias-adjusted estimator (44) reduces to

$$\hat{\mathbf{Y}}_{I(d)}^a = \hat{p}^{-1} \hat{\mathbf{Y}}_{I(d)} + (1 - \hat{p}^{-1}) \hat{\mathbf{T}}^{-1} \sum_s w_i \mathbf{x}_i \mathbf{z}_i' \hat{\gamma}_r, \quad (45)$$

noting that $\hat{\gamma}_r = \hat{\gamma}_I$, where, under deterministic regression imputation,

$$\begin{aligned} \hat{\gamma}_I &= \left(\sum_{i \in s} w_i \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i) \right)^{-1} \\ &\times \left[\sum_{i \in s} w_i a_i \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i) + \sum_{i \in s} w_i (1 - a_i) \mathbf{z}_i y_i^* / (\lambda' \mathbf{z}_i) \right] \\ &= \hat{\gamma}_r. \end{aligned}$$

Haziza and Rao (2005) obtained the bias-adjusted estimator (45).

Concluding Remarks

For simplicity, we focussed on a single imputation class but our GNM method readily extends to multiple imputation classes by using separate imputations across classes. For example, we could use weighted mean imputation within classes using our modified weights \tilde{w}_i . Also, our method can be extended to the case of composite imputation (Sitter and Rao 1997; Shao and Steel 1999) which uses different imputations for missing item values depending on the auxiliary information available. For example, ratio imputation is used when an auxiliary variable x is observed and some other imputation when x is not observed. In this case, the IM approach based on the ratio model relating y to x will not be applicable unlike in the case where x is observed on all the sampled units.

Acknowledgments

J.N.K. Rao's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors wish to thank the reviewers for useful comments and suggestions.

References

- Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society*, B, 67, 445-458.
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 15, 279-292.
- Brick, J.M., and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.
- Deville, J.C., and Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.
- Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, 429-440.

- Haziza, D., and Rao, J.N.K. (2005). Inference for domains under imputation for missing survey data. *Canadian Journal of Statistics*, 33, 149-161.
- Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 2, 169-174.
- Rao, J.N.K. (1990). Variance estimation under imputation for missing data. Technical report, Statistics Canada, Ottawa.
- Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of American Statistical Association*, 91, 499-506.
- Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31, 117-138.
- Rao, J.N.K., and Shao, J. (1992). On variance estimation under imputation for missing data. *Biometrika*, 79, 811-822.
- Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
- Särndal, C.-E. (1992). Method for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.
- Shao, J., and Steel, P. (1999). Variance Estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.
- Sitter, R., and Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *Canadian Journal of Statistics*, 25, 61-73.

A Model for Estimating and Imputing Nonrespondent Census Households under Sampling for Nonresponse Follow-up

Elaine L. Zanutto and Alan M. Zaslavsky¹

Abstract

Sampling for nonresponse follow-up (NRFU) was an innovation for U.S. Decennial Census methodology considered for the year 2000. Sampling for NRFU involves sending field enumerators to only a sample of the housing units that did not respond to the initial mailed questionnaire, thereby reducing costs but creating a major small-area estimation problem. We propose a model to impute the characteristics of the housing units that did not respond to the mailed questionnaire, to benefit from the large cost savings of NRFU sampling while still attaining acceptable levels of accuracy for small areas. Our strategy is to model household characteristics using low-dimensional covariates at detailed levels of geography and more detailed covariates at larger levels of geography. To do this, households are first classified into a small number of types. A hierarchical loglinear model then estimates the distribution of household types among the nonsample nonrespondent households in each block. This distribution depends on the characteristics of mailback respondents in the same block and sampled nonrespondents in nearby blocks. Nonsample nonrespondent households can then be imputed according to this estimated household type distribution. We evaluate the performance of our loglinear model through simulation. Results show that, when compared to estimates from alternative models, our loglinear model produces estimates with much smaller MSE in many cases and estimates with approximately the same size MSE in most other cases. Although sampling for NRFU was not used in the 2000 census, our estimation and imputation strategy can be used in any census or survey using sampling for NRFU where units are clustered such that the characteristics of nonrespondents are related to the characteristics of respondents in the same area and also related to the characteristics of sampled nonrespondents in nearby areas.

Key Words: Missing data; Small area estimation; Iterative proportional fitting; Log-linear models; ECM.

1. Introduction

Sampling for nonresponse follow-up (NRFU) was an innovation for U.S. Decennial Census methodology considered for the year 2000 (U.S. Bureau of the Census 1997a, b). Under current procedures used in 99% of households, the Census Bureau first mails or personally delivers a questionnaire, to be returned by mail. Then field enumerators attempt to contact all mail nonrespondents (about 35% of those mailed). The workload of about 42 million households makes this one of the most expensive census operations.

Sampling for NRFU involves sending field enumerators to only a sample of the nonresponding housing units. This sample is either an unclustered element sample of nonresponding housing units (the "unit sample") or a cluster sample consisting of all nonresponding units in a sample of the census blocks (small areas approximating a city block or some compact rural area, averaging about 15 housing units). This second stage of followup leads to the completion of a questionnaire (through proxy response or imputation, if necessary) for all sample housing units, except those that are resolved to be vacant.

The potential cost savings of sampling are large, but it would require estimating the characteristics of a huge

number of nonsampled nonresponding households, posing a major small-area estimation problem (Ghosh and Rao 1994; Rao 2003). We show that using appropriate models to impute the characteristics of the nonsample nonrespondent households, we may benefit from the large cost savings of NRFU sampling while still attaining acceptable levels of accuracy for small areas. Our strategy is to model household characteristics using low-dimensional covariates at detailed levels of geography and more detailed covariates at larger levels of geography. To do this, households are first classified into a small number of types. A hierarchical loglinear model then estimates the distribution of household types among the nonsample nonrespondent households in each block. This distribution depends on the characteristics of mailback respondents in the same block and sampled nonrespondents in nearby blocks. Nonsample nonrespondent households can then be imputed according to this estimated household type distribution.

Although, for complex legal reasons, sampling for NRFU was not used in the 2000 census, our estimation and imputation strategy can be used for small area estimation or imputation in any census or survey using sampling for NRFU where units are clustered such that the characteristics of nonrespondents are related to the characteristics of

1. Elaine L. Zanutto, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, U.S.A. E-mail: zanutto@wharton.upenn.edu; Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A. E-mail: zaslavsky@hcp.med.harvard.edu.

respondents in the same area and also related to the characteristics of sampled nonrespondents in nearby areas. The related methodologies of Purcell and Kish (1980) and Zhang and Chambers (2004) also use loglinear models to estimate small-area cross-classified counts assuming that the total populations are known and that auxiliary cross-classified data is available at the small area level. We have an additional source of information, specifically the characteristics of the nonrespondents in the NRFU sample. This allows us to model the relationship between respondents and nonrespondents directly in some blocks.

Section 2 summarizes proposed strategies for imputing missing data in this situation. Section 3.1 describes our general sampling and estimation procedure. We present our estimation and imputation model in Section 3.2, our smoothing and estimation procedures in Section 3.3, and evaluate our model by simulation in Section 4. Methods for MSE estimation are summarized in Section 5, and Section 6 presents conclusions.

2. Previous Proposals for Imputing Census Nonrespondents

Several methods have been proposed for imputing the characteristics of nonresponding housing units. "Top-down" strategies first estimate counts for aggregates of households and then allocate them to small areas in a manner that maintains consistency with the aggregates. Simple ratio models (Fuller, Isaki and Tsay 1994, henceforth "FIT"), Poisson regression models (Bell and Otto 1994), or more complex loglinear models (as we propose here and in Zanutto and Zaslavsky 1995b, a) are used to estimate counts for small areas and detailed demographic groups for which direct estimates are not possible. Like us, FIT classify households into a modest number of types defined by important characteristics (e.g., number of people, race, tenure) and then estimate the number of households of each type among nonsample nonrespondents. A complete census roster is then generated by imputing the estimated number of households of each type. The main difference between our approach and that of FIT is that by using a loglinear model rather than a stratified ratio model, we obtain more flexibility in the detail of constraints imposed at various levels of geography. Bell and Otto (1994) estimate the number of people over 18 years old of each race (Hispanic, non-Hispanic Black, Other) in each nonsample nonrespondent housing unit but do not consider how to group imputed persons into households or how to impute household-level characteristics such as tenure. These *ad hoc* "top-down" models incorporate at most a few household characteristics and hence do not explicitly model household structure, but they are designed to maintain the consistency of the aggregates that are considered most important.

Schafer (1995) develops a "bottom-up" strategy in which households are built up from individual persons and their characteristics and relationships, each of which must be described by its own model. These models describe the population in more detail and can support full probability (e.g., Bayesian) inferences about unobserved characteristics. However, this approach, unlike the other, requires that a fairly complex set of models be built before any imputations can be made. Furthermore, in this framework it is more difficult to maintain consistency between microdata and aggregate controls. A combined strategy, however, could use our models to produce nearly unbiased estimates by household types and Schafer's models to complete the imputations.

3. Estimation Procedures and Models

3.1 Overview

In the first step of the imputation procedure, counts of the number of nonsample nonrespondent households of each type are predicted using a combination of logistic and loglinear models for each block. This step is the topic of this paper (and of FIT).

For modeling we classified households into types based on a few important characteristics. Here we use 19 types, one of which is "vacant." The remaining 18 are defined by the cross-classification of households by three size categories (1–2 people, 3–4 people, 5 or more people), three race categories (Hispanic, non-Hispanic Black, Other), and two tenure categories (owner, renter).

To predict the number of vacant housing units among nonsample nonresponding units in each block we (and FIT) fitted a logistic regression model, recognizing that the relationship between respondent and nonrespondent households is different for vacant than for nonvacant housing units. Respondent vacants are simply those that were identified as vacant by a postal service letter carrier, leading to mail return of the original questionnaire. Their distribution is likely to depend largely on housing characteristics related to postal delivery, telling us little about the distribution of nonrespondent vacants.

After modeling vacancies, we fitted a loglinear model to predict the distribution of the nonvacant household types in the remaining nonsample nonrespondent households at three geographical levels. The block is the smallest unit and the one for which estimated counts are calculated. The "estimation domain" is the largest unit and is the area in which estimation is conducted independent of other such domains; in our application to the 1990 census, this is the area for which the census was administered from one of 449 district field offices (DO) representing about 200,000

households on average. Finally, we call an intermediate level of geography an “area”, comprising a relatively homogeneous collection of contiguous blocks within an estimation domain. In standard Census Bureau geography these might be census tracts, block groups, or Address Register Areas.

We lay out briefly the remaining steps that would be followed to obtain census products using the estimates. In the second step of the imputation procedure the predicted counts would be rounded to integers. Unbiased schemes (*i.e.*, stochastic procedures that in expectation impute the predicted number of units in each cell) for “controlled rounding” (*i.e.*, rounding in a two-way table while preserving marginal totals) were developed by Cox (1987) and George and Penny (1987). However, more research is needed to determine if these methods can be modified to round households counts while preserving all the margins corresponding to effects in the loglinear model. This is an active research topic due to its importance to statistical nondisclosure.

Finally, detailed person and household information would be imputed for nonrespondent households by substituting donor households with similar characteristics. Donors can be chosen from the sampled nonrespondents, the respondents, or a combination of both sources. Finally, tabulations and microdata samples would be prepared from the completed rosters.

3.2 Loglinear Model

We fitted a loglinear model to estimate the prevalence of the various types of households among nonsample nonrespondent households in a DO, using data from the respondents and from the nonrespondents in the NRFU sample for that DO. The model predicts household types for nonsample nonrespondent households in each block by using information about the characteristics of respondent households in the same block and the characteristics of nonrespondent households, measured by the NRFU sample, in surrounding blocks. To accomplish this, the loglinear model contains interactions among the household characteristics that define household type and response status at various levels of geography.

This modeling strategy is motivated by the fact that when a hierarchical loglinear model (*i.e.*, one in which for every included interaction effect, all main effects or interactions marginal to it are also included) is fitted by maximum likelihood, the fitted values for every margin or mean corresponding to an effect in the model are equal to the corresponding observed margins or means (Birch 1963). Therefore, predictions for household types agree with observed rates for the characteristics included in the model, at the levels of geography and response status corresponding

to the interactions included in the model. Also, because model predictions for the included effects are constrained to agree with observed rates based on a probability sample (the NRFU sample), the corresponding estimates are consistent and approximately unbiased. (Exact unbiasedness is not obtained because of the nonlinearity of the prediction model and because the number of nonsample nonrespondent households in a block might be associated with some characteristics of the nonresponding households in the block.)

The loglinear model includes nested geographical factors for blocks and areas. It also includes crossed factors representing the demographic characteristics of households: first-stage response indicator (respondent or nonrespondent household), household type index, and model expressions in the variables that define household types. These model expressions are submodels of the fully interacted model which defines household type (*i.e.*, race \times size \times tenure).

We use the following notation:

- i = block index ($i = 1, \dots$, number of blocks in the DO),
- j = index of household type ($j = 1, \dots$, number of types),
- r = first-stage (mail) response indicator,
 $r = 0$ for nonresponding households
and $r = 1$ for respondents,
- $a = a(i)$ = index for the area containing block
 i ($a = 1, \dots$, number of areas),
- $x_k = x_k(j)$ = model expressions in the variables that
 $k = 1, 2, 3, 4$ define household types where x_1 represents the full cross-classification defining household types, x_2 and x_3 are model expressions which are marginal to x_1 , and x_4 is a model expression which is marginal to x_3 . (This terminology is explained below.)

We assume a loglinear model of the following form:

$$n_{ijr} \sim \text{Poisson}(m_{ijr}), \log(m_{ijr}) = z_{ijr}^T \beta \quad (1)$$

where n_{ijr} and m_{ijr} are respectively the observed and expected counts for block i , household type j and response status r , and Z is the design matrix corresponding to the following model formula:

$$x_1 + i * x_2 + i * r + r * x_3 + r * a * x_4. \quad (2)$$

In the standard generalized linear models notation of Wilkinson and Rogers (1973), the “*” operator indicates that the main effects and all interactions that are marginal to the given interaction are included in the model, so that this model contains main effects for model expression x_1 , response indicator r , and block indices i and the interactions $i * x_2$, $i * r$, $r * x_3$, and $r * a * x_4$.

Because, in (1), x_4 interacts with area, the smallest level of aggregation for the non-respondent data, it should represent a fairly coarse classification of households including only those household characteristics that are most important to impute accurately at the area level. The x_3 expression may include terms not included in x_4 , since it is fitted at a higher level of geography where there is more data available. Similarly, the x_1 expression might include the most interactions, including the interaction of all variables that define household type, since it is fitted at the largest level of geography, using all available data. Finally, x_2 , which can be different than x_3 since it interacts with i instead of r , should be less detailed than x_1 since it interacts with block, a much smaller level of geography. These guidelines are motivated by the fact that estimates of interactions with i , r , or a are determined by relatively few observations and should be kept simple. Choosing x_2 , x_3 , and x_4 as described above should improve the precision of model estimates while preserving the most important margins.

As an example of possible x_1, \dots, x_4 terms, suppose that we define household type by a race \times size \times tenure cross-classification. Then one possible specification of x_1 , x_2 and x_3 is $x_1 = \text{race} * \text{size} * \text{tenure}$, $x_2 = \text{race} * \text{size} + \text{tenure}$, $x_3 = \text{size} * \text{tenure}$, and $x_4 = \text{race} + \text{size} + \text{tenure}$. Allowing the x_1, \dots, x_4 terms to be model expressions, rather than just simple interactions, gives us a concise way to represent a model containing all the desired interactions. For example, a model containing an $i * x_2$ term, where x_2 is specified above, includes both a block \times race \times size interaction and a block \times tenure interaction.

A heuristic interpretation of our loglinear model is that we estimate the detailed distribution of household types across the whole area (x_1) and then shift that distribution to allow for the general characteristics of the block (x_2), the general differences between responding and nonresponding households (x_3), and the most important differences between responding and nonresponding households in the particular area (x_4). All interactions could be included except those of the form $r * i * x$, where x represents a model expression in the variables that define household type (*i.e.*, such as x_1 , x_2 , x_3 , or x_4). Interactions of this form depend on the margins determined only by non-respondent households in a single block and these are unavailable in nonsample blocks under the block sample design, and based on a very small sample under the unit sampling design. Therefore our model specification excludes all $r * i * x$ effects, which are always inestimable (or poorly estimated, in the household sample design). This model generalizes two simple theories which are contained as submodels. First, if there are no differences between blocks (*i.e.*, the loglinear $i * x_2$ and $a * x_4$ interactions are zero) then

nonrespondent households in each block are imputed according to the overall proportion of nonrespondent households in each of the x_3 categories in the NRFU sample, through the $r * x_3$ effect. In other words, the imputations are made using the same proportions in each block. Second, if there are no differences between respondents and nonrespondents (*i.e.*, no $r * x_3$ or $r * x_4$ interactions) then nonrespondents are imputed in the same proportions as observed in the respondents in each block.

Our general model formulation can accommodate many definitions of area and household type and choices of model expressions. Areas should be defined to be large enough to contain adequate data to estimate the corresponding interactions, but also relatively homogeneous. For example, areas could be defined by a combination of geographical contiguity and stratification by block-level covariates (such as percent minority), in order to obtain more homogeneous areas whose differences could be described by modeling. Generalization to more than two levels of geography within the estimation domain is also straightforward. Thus, for example, we could interact another model expression x_5 with a geographical unit intermediate between the area and the block.

Fitting the model by maximum likelihood, the following quantities are made equal to the corresponding observed values: (1) fitted block counts (through the main effect for block, i), (2) response rates by block (through the $r * i$ term), (3) household characteristic means overall (for x_1 characteristics through the main effect term for x_1) and (4) by block (for x_2 characteristics through the $i * x_2$ term), and (5) household characteristic means for nonrespondents overall (for x_3 characteristics, through the $r * x_3$ term) and (6) for nonrespondents by area (for x_4 characteristics, through the $r * a * x_4$ term). Thus, this model generalizes the model used by FIT of block \times type independence, yielding unbiasedness at smaller levels of aggregation, assuming that the margins and averages are estimated unbiasedly from the data. The estimate for area is not exactly the same as the usual unbiased estimate obtained by direct estimation from the NRFU sample because the model makes observed and fitted margins agree for the households in sample. In effect, there is covariance (regression) adjustment that shifts the aggregate to account for observed differences between respondent households in sample blocks and respondent households in nonsample blocks, or in the unit sampling design, between respondent households in blocks with households in the NRFU sample and blocks without households in the NRFU sample.

The idea of modeling household characteristics using low-dimensional covariates at the block level and in more detail at more aggregated levels is similar in concept, although not in details, to the model described in Zaslavsky

(2004). For use of loglinear weights to match sample estimates of aggregates, see Brackstone and Rao (1976), Oh and Scheuren (1983), and Zaslavsky (1988).

3.3 Estimation and Smoothing

We fit the model by maximum likelihood estimation under the Poisson sampling model, which is equivalent to fitting a multinomial logistic regression model. The fitting is complicated by the fact that the data do not form a complete block \times response \times type table because we have counts by block, but not characteristics for nonsample nonresponding households. In the block sampling design we lack characteristics of all nonrespondents in some blocks and in the unit sampling design we lack characteristics of some nonrespondents in almost all blocks. To fit the model we use a modified iterative proportional fitting (IPF) algorithm adapted to data that are partially classified in a part of the dataset (Appendix).

With some data sets, some parameters may be inestimable because the maximum likelihood estimates lie on the boundary of the parameter space (infinite on the loglinear scale, indicating a zero on the count scale) or because there is no information for the parameter. Tailoring the model specification in each estimation domain to remove inestimable parameters is impractical in a census production setting.

By introducing a small amount of prior information, estimability of all parameters can be guaranteed. To do this, we append a small amount of "pseudo-data" to the data for each area, whose proportions by type are equal to those for some surrounding area (the DO, in our simulations), by adding these counts to the data table before fitting the model. This implements an empirical Bayes analysis for multinomial data with distribution $f(n_1, \dots, n_H | p_1, \dots, p_H) \propto \prod_{i=1}^H p_i^{n_i}$, where n_1, \dots, n_H are the observed number of households of each type in a block or area. If $\{p_i\}$ have a joint Dirichlet prior distribution, $f(p_1, \dots, p_H) \propto \prod_{i=1}^H p_i^{\alpha_i - 1}$, $\alpha_i \geq 0$, the resulting posterior distribution for the p_i 's is Dirichlet with parameters $\alpha_i + x_i$ (Gelman, Carlin, Stern and Rubin 1995, page 76) and posterior mode proportional to the parameters. Thus, this empirical Bayes procedure is equivalent to adding $\sum \alpha_i$ households to the area, where α_i of these households are of the i^{th} type. We fix the α_i 's to be proportional to the observed proportions of each household type in some surrounding area, thus avoiding introducing bias at the level of the larger area. This prior specification induces a prior on the parameters of the loglinear model. See Rubin and Schenker (1987), Zaslavsky (1988), and example and

historical references in Clogg, Rubin, Schenker, Schultz and Weidman (1991) for similar use of smoothing.

After estimating the model parameters, the next step is to calculate predicted counts for each household type for the nonrespondent households that are not in the NRFU sample. Using the IPF algorithm, the predictions for the nonsample nonrespondent households are obtained automatically by applying the same fitting proportions to the partially observed part of the table as to the fully observed part of the table, so no further calculation is required (Appendix).

4. Simulations

4.1 Overview

Our simulation study evaluated the bias, variance and MSE of the estimates of estimated demographic aggregates (such as the number of households by race, size and tenure) at various levels of geography, using estimated household compositions for non-respondent households that are not in the NRFU sample. Analytic evaluations are infeasible, given the complexity of the models and sampling scheme, the dependence of the performance of the model on the actual geographical distribution of household types, and the number of variations of the model that could be examined.

We used block-level data from three DOs from the 1990 U.S. Decennial Census; these constituted our estimation areas. The simulations are similar in structure to those described by Schindler (1993) or FIT.

The steps of the simulation are as follows:

1. Blocks or nonrespondent housing units are sampled according to the NRFU sampling scheme.
2. A logistic regression model for vacant households is fitted to the respondent households and the sampled nonrespondent households.
3. The predicted number of nonrespondent households that are vacant is calculated for each block.
4. A model for nonvacant types is fitted using the respondent households and the sampled non-respondent households.
5. The predicted number of nonsample nonrespondent households of each nonvacant type are calculated for each block.
6. Aggregates of interest are calculated based on the predicted counts, and compared to the truth using loss functions.

In our simulations, repeating these steps 30 times yielded estimates of RMSE (defined in section 4.3) with adequate accuracy to evaluate the performance of our model relative to the alternative models. Specifically, the estimated coefficients of variation of the estimated differences in RMSE for the stratified ratio method (described below) and

loglinear method are less than 0.05, except when the difference between estimated RMSEs is very small, resulting in a large coefficient of variation.

The performance of our proposed model is compared with two alternative estimation methods, under both the unit and block sampling designs. Each method first fits a logistic regression model to estimate the number of nonrespondent households that are vacant in each block. The first alternative, the “unstratified ratio method”, imputes households for nonsample nonrespondent households in each block in proportion to the distribution of household types among nonrespondent households in the follow-up sample for the entire DO. The second alternative, the “stratified ratio method”, is a version of that in FIT. We first form strata of approximately 82 blocks based on the racial composition of the blocks, as described by FIT. (We use both respondent and nonrespondent data to form strata, assuming, as in FIT, that similar information would be available from administrative records. Stratification based only on respondent information yielded similar results.) Then, in each stratum, nonsample nonrespondent households are imputed to non-vacant types in proportion to the frequency of the type in the follow-up sample for that stratum.

We simulate each estimation method using a NRFU sampling rate of 30%. In each stratum, we simulate NRFU sampling by selecting a 30% simple random sample of blocks for the block sampling design, and a 30% simple random sample of nonrespondent households in each stratum for the unit sampling design. The characteristics of the nonrespondent households in these samples is assumed to be known (*i.e.*, as a result of follow-up operations). For both our loglinear model method and the stratified ratio method, we select a 30% sample of blocks or nonrespondent households using simple random sampling without replacement from each area.

We considered several loglinear model formulations. The best model for both the block and unit sampling designs, by the criteria described in Section 4.3, uses $x_1 = \text{size} * \text{race} * \text{tenure}$, $x_2 = \text{race} * \text{tenure} + \text{size}$, $x_3 = \text{race} * \text{size}$, $x_4 = \text{tenure}$. This model is used in the simulations.

To ensure the model can be fitted in every case and to speed the convergence of the IPF, we smooth the data by adding one hypothetical respondent household (“pseudo-data”) to each block. This household is divided among the 18 nonvacant household types according to the overall DO proportions of respondent households. Estimates using 5 households for smoothing were about as accurate as with one, and more aggressive smoothing (adding 10, 15, 20 or 25 households per block) slightly increases errors in the estimates. Also, although adding only a small fraction of a household to each block is sufficient to ensure that the model can be fitted in every case, using less than 1

household per block drastically slowed convergence and slightly increased the error in the estimates.

The three estimation procedures used the same logistic regression model for vacancies. The covariates for each block are the mail nonresponse rate, the percentages of respondent households that are (separately) renters, apartment dwellers, and of a minority race (either Black or Hispanic), the average value of owner-occupied homes, the average monthly rent for rental units, indicator variables for each of the areas, and interactions between percentage of respondent renters and average monthly rent, percentage of respondent renters and average monthly rent squared (mean-centered), percentage of respondent owners and average home values, and percentage of respondent owners and average home values squared (mean-centered). To avoid computational problems arising from blocks with no nonrespondent vacant households, one hypothetical nonrespondent household is added to each block divided between vacant and nonvacant according to their proportions in the sampled nonrespondent households in the DO.

4.2 Data

We use short-form data from the 1990 census for three DOs, whose characteristics are described in Table 1. The race of a household is determined by the most prevalent race in the household, usually (98% of households) the only race. In DO 1 we grouped consecutive (and therefore contiguous) block groups (clusters of contiguous blocks) into 94 areas containing an average of 52 blocks and 1100 households. For DOs 2 and 3, block group information was unavailable so we formed areas by grouping consecutive blocks into clusters containing an average of 50 blocks (on average, 548 households per area in DO 2 and 918 households per area in DO 3).

Table 1
Characteristics of the Census District Office Areas
Used in the Simulations

	DO1	DO2	DO3
Household	112,966	169,321	149,567
Blocks	4,907	15,470	8,167
Pseudo-areas	94	309	163
Non-Hispanic Black	14.4%	28.5%	1.3%
Hispanic	6.1%	1.0%	6.6%
Other	73.5%	59.4%	81.5%
Owner	63.8%	59.5%	52.6%
Renter	30.2%	29.4%	36.7%
Vacant	6.0%	11.1%	10.7%
Size 1 (1–2 people)	50.4%	46.9%	55.2%
Size 2 (3–4 people)	31.6%	31.6%	26.2%
Size 3 (5+ people)	12.0%	10.4%	7.9%
Response Rate	72.6%	65.3%	56.7%

4.3 Measures of Bias, Variance, and Mean Squared Error

Loss functions for our evaluations are based on the relative error for household category j (a type or combination of types) in geographic area i (a block or collection of blocks):

$$d_{ijs} = \frac{\hat{Y}_{ijs} - Y_{ij}}{Y_{i+}} \quad (3)$$

where Y_{ij} is the true number of households of category j in geographical unit i , \hat{Y}_{ijs} is the corresponding number of households estimated from sample s (including those observed in the sample and estimated by the model), and Y_{i+} is the total number of households in geographical unit i .

We summarize bias in estimated counts for category j and a level of geography (block, area, DO) with Root Mean Weighted Squared Bias (RMWSB):

$$\hat{\text{RMWSB}}_j^2 = \frac{\sum_i Y_{i+} \left\{ \left(\frac{1}{S} \sum_s d_{ijs} \right)^2 - \frac{1}{S(S-1)} \left(\sum_s d_{ijs}^2 - \frac{1}{S} \left(\sum_s d_{ijs} \right)^2 \right) \right\}}{\sum_i Y_{i+}} \quad (4)$$

where S is the number of samples drawn and $i=1, \dots, I$ where I is the number of geographical units. The second term in the numerator removes a bias due to the finiteness of the simulation. From a design-based perspective, we regard the composition of each area as a fixed quantity, and only sampling is random. Then bias is defined as the average difference, over all possible samples, between the truth for an area and the corresponding estimates, essentially the model error for that area. Such error is inevitable since the composition of the nonrespondents in any block is not entirely predictable. A more serious type of bias would involve systematic error in estimates for a collection of blocks with similar composition. Although we have not checked for all possible types of bias in this sense, the model specification protects us against bias at higher levels of aggregation because model estimates are constrained to agree (approximately) with unbiased estimates for areas and DOs.

As a measure of overall error, we calculate the Root Mean Weighted Mean Squared Error (RMWMSE) for each household category j , which is given by

$$\hat{\text{RMWMSE}}_j^2 = \frac{\sum_i Y_{i+} \left(\frac{1}{S} \sum_s d_{ijs}^2 \right)}{\sum_i Y_{i+}} \quad (5)$$

where Y_{ij} , \hat{Y}_{ijs} , Y_{i+} , i , and S are defined as above. (The two “means” refer to mean over geographical units (i) and over samples (s).) We obtain a measure of the standard deviation of the estimates for household category j by calculating the Root Mean Weighted Variance (RMWV):

$$\begin{aligned} \hat{\text{RMWV}}_j^2 &= \frac{\sum_i Y_{i+} \left\{ \frac{1}{S-1} \left(\sum_s d_{ijs}^2 - \frac{1}{S} \left(\sum_s d_{ijs} \right)^2 \right) \right\}}{\sum_i Y_{i+}} \\ &= \hat{\text{RMWMSE}}_j^2 - \hat{\text{RMWSB}}_j^2. \end{aligned} \quad (6)$$

Note that these MSE, bias, and standard deviation measures are all estimates of expectations with respect to repeated NRFU sampling from the given finite population of blocks. These loss functions can be applied at various levels of geography, reflecting the fact that the main use of block level estimates is aggregation to form estimates at higher levels of geography. With this in mind, these measures were also chosen because they weight errors by the size of the geographical unit. This leads to consistent estimates of error when aggregating over geographical units, which is appropriate due to the arbitrariness of unit boundaries (Zaslavsky 1993). We base our measures on errors relative to the total area i population rather than the population in the target category only, because the latter denominator inflates the importance of small errors in blocks where the category rarely or never appears.

4.4 Results

For simulated NRFU sampling using both the block and unit sampling designs, estimates of the number of households with each characteristic are calculated at block, area, and DO levels of geography using each of the three estimation methods. The results for each method are represented by the shaded bars in Figure 1 for the unit sampling design. (Results for the block sampling design are not shown here, but the pattern of results are similar with the RMWMSE being about 10% greater for all estimates.) In this figure, each row of bar charts displays the RMWMSE for block, area, and DO level estimates for one of the three DOs. Each group of three bars represents the RMWMSE for estimates of the total number of households for each of the tenure categories, the household size categories and the race categories using each of the three methods. Because all three methods use the same logistic regression model to predict the number of vacant nonsample nonrespondents in each block, the vacant category is omitted from the plots.

RMWMSE with both the stratified ratio method and the loglinear model was much smaller than with the unstratified ratio method for most household characteristics at the block and area level. Therefore, we confine further discussion to comparison of the two former methods.

The most dramatic differences appear for the tenure categories at the block and area levels. In each DO, block and area level estimates of the tenure categories from the loglinear model have much smaller RMWSE than the estimates from the stratified ratio method, primarily because the former had much smaller bias (RMWSB). Standard deviations (RMWV) were slightly larger for the loglinear model under the unit sampling design, but about equal for the two methods under the block sampling design. The loglinear model had smaller bias for the tenure categories at the area level because tenure is included in the model as an area-level effect, x_4 . Stratification on race in the ratio method reduces RMWSE for the race categories at the block level, but the two methods have comparable

RMWSE for the race categories at the area and DO levels. The stratified ratio method loses its advantage over the loglinear model at the area level because the former does not use any area-level information. Both methods generally produce estimates with comparable RMWSE at all levels of geography for the size categories.

The statistical significance (under the simulations) of differences in RMWSE between the methods was evaluated using t -tests. Almost all differences at the block and area levels, excluding the vacant category, have two-tailed p -values ≤ 0.001 and therefore cannot be attributed to simulation error.

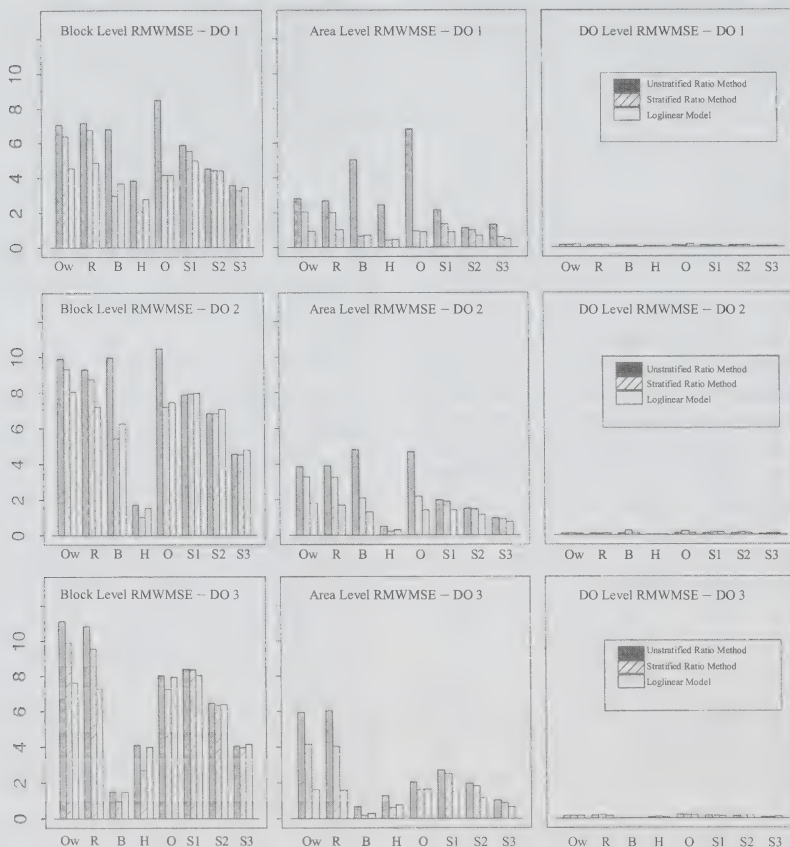


Figure 1. RMWSE for block, area, and DO level estimates for each household characteristic, using the unit sampling design for DO 1, 2, and 3, with 30 simulated samples ("Ow" = Owner, "R" = Renter, "B" = Black, "H" = Hispanic, "O" = Other race, "S1" = Size group 1 (1–2 people), "S2" = Size group 2 (3–4 people), "S3" = Size group 3 (5 or more people)).

5. Assessment and Prediction of Model Error

Methods for estimation of MSE of fitted estimates using sample data are briefly summarized here due to space limitation; methods and findings are available from the first author.

First, we developed analytic approximations that predict the effect of changing the sampling rate on the accuracy of our estimates without requiring additional simulations at each rate. These can be useful for sample design. We approximate the RMWSE of block, area, and DO level estimates at a new sampling rate under both the block and unit sampling designs, assuming simulation results using one sampling rate are already available, by combining estimates of bias and variance at the current sampling rate using two rescaling factors. The first factor reflects the changed proportion of housing units that require estimation under the new sampling rate, which affects the bias and variance of the combined estimates. The other reflects the effect of the sampling rate on the variance of the estimates for the nonresponding units. Simulations demonstrated the accuracy of predictions for RMWSE using these approximations, except for some extreme extrapolations.

Using these results, we developed a cross-validation procedure to facilitate within-sample estimates of RMWSE for use in a production setting where the true characteristics of the nonsample nonrespondent households are not known. The follow-up sample in each area is divided randomly into C cross-validation groups (of blocks for block sampling, and of households for unit sampling). Each cross-validation group is dropped out in turn and the model is fitted to the nonrespondents in the remaining $C - 1$ cross-validation groups and the respondents in all C groups. We can then estimate RMWSE under the design simulated by the cross-validation and project this estimate to the actual sampling rate, or some other rate of interest, using the approximations described in the preceding paragraph. Simulations show that this produces accurate estimates of RMWSE at block and DO levels of geography, with some overestimation at the area level. This method also provides separate estimates of bias and variance that are shown by simulation to be very accurate. These are useful for assessing model adequacy since a poorly-fitting model would be betrayed by a large component of MSE due to bias.

6. Conclusions

In the preceding sections, we have presented a model-based approach to imputation of the characteristics of nonresponding households in a census that were not sampled for nonresponse followup. In simulations, our loglinear model produces estimates with much smaller error

than two alternatives for some estimands, and is about equivalent for others. These conclusions hold for both the block and unit sampling designs. An advantage of our approach is that models can be specified to constrain only a few marginal tables or interactions of characteristics at the finest levels of geography, where the data are sparse, while fitting more detailed distributions of characteristics at higher levels of geographic aggregation at which more data are available. This is consistent with typical practice in release of census data, which include minimal characteristics at the block level but increasingly more detailed characteristics for larger units.

Many important uses of the census involve estimation of the population and its characteristics for small domains such as legislative districts and planning areas for social services (such as schools and clinics) and commercial development. Even though these domains will not always align with the areas used in census estimation, controlling the census estimates to match unbiased estimates at several levels of geography makes it more likely that estimates for policy-relevant domains assembled from wholes or parts of these areas will also be nearly unbiased. Our method has more predictable aggregate properties than complex alternatives such as hierarchical spatial modeling. Although the latter might produce estimates with smaller MSE at the lowest levels of geography, fitting such models and checking their biases at various levels of geographic aggregation would require extensive local tuning which is likely to be impractical in a census production setting.

Our methodology is illustrated here in the context of a NRFU sampling for the U.S. Decennial Census, but our estimation and imputation strategy can be used for small area estimation or imputation in any census or survey using sampling for nonresponse followup with hierarchically structured populations. We can also incorporate administrative records as covariates for predicting the characteristics of the corresponding nonrespondent households (Zanutto and Zaslavsky 2002). In that scenario, data from households in the NRFU sample for which we have both census and administrative records information are used to estimate the systematic differences between the two information sources. Under the same models, we impute the characteristics of nonsample nonrespondent households. Using administrative records through this modeling approach can improve the accuracy of small area (block-level) estimates.

Although the discussion of sampling in the United States census has been politically contentious, nonetheless in the long run it seems likely that some form of estimation will be used for nonrespondents. The potential might be even greater in countries where population estimation already makes substantial use of administrative records (Redfern 1989). Methods such as those described here that can

combine information across data sources while reflecting local diversity will be essential to such efforts.

Appendix

Iterative Proportional Fitting with Partially Cross-Classified Data

A standard approach to fitting loglinear models to partially cross-classified data uses an EM algorithm (Dempster, Laird and Rubin 1977; Little and Rubin 2002, chapter 8), in which in alternate steps (1) the expected counts are imputed under the model and (2) the model is refitted to the observed and imputed data, using iterative proportional fitting (IPF) (Darroch and Ratcliff 1972) for models without closed-form solutions. In the more efficient ECM modification of this algorithm, only a single cycle of the IPF algorithm is taken at each step (Meng and Rubin 1993).

For our application we developed a modified IPF algorithm that is faster than the EM and ECM algorithms for our models, which always include a block \times response interaction and never include any block \times type \times response interactions. We found that our modified IPF algorithm converges in approximately one half to two thirds the number of cycles that ECM requires with less computation per step (Zanutto 1998, Part 1, Appendix A). (Convergence is declared when the predicted and observed values of the minimal sufficient statistics of the model are sufficiently close.)

Our algorithm takes advantage of the fact that partially classified observations contribute to the likelihood only through the total number of nonrespondent households in each block. Therefore, to maximize this part of the likelihood we need only ensure that the fitted number of nonrespondents in each block equals the observed number, which is automatic because the block \times response interaction is always included in our model.

The modified IPF algorithm fits the model to the fully classified observations using an ordinary IPF algorithm, ignoring the partially cross-classified observations. For the block sampling design, this means that the model is fitted using the fully observed part of the block \times type \times response table using an ordinary IPF algorithm, ignoring the partially classified part of the table. Predictions for the partially cross-classified cells are obtained by applying the same fitting proportions to those cells as to the fully observed part of the table. Finally, predictions for the partially cross-classified cells are scaled so that the fitted number of nonrespondents in each block equals the observed number. For the unit sampling design, the same algorithm is used, viewing the collection of respondent households and nonrespondent households in the follow-up sample as analogous to the fully-observed part of the table in the block

sampling design and viewing the blocks with no nonrespondents in the follow-up sample as analogous to the out-of-sample blocks in the block sampling design. This gives predictions for nonrespondent households in blocks with no nonrespondents in the follow-up sample. Predictions for nonrespondent households in blocks with one or more nonrespondent households in the follow-up sample are obtained by applying the predicted distribution of household types among sampled nonrespondent households in each of these blocks to the corresponding nonsample nonrespondent households in these blocks. For more details about in the unit sampling case, see Zanutto and Zaslavsky (2002).

We now illustrate the IPF algorithm for the block sampling design under a Poisson model like (1) with $\log(m_{ijr}) = z_{ijr}^T \beta$ where m_{ijr} represents the expected number of households in block i of household type j of response status r , and Z is the design matrix corresponding to the model expression $i * x + i * r + r * x$. This is a simplified version of the model in (2) with only one level of geography and only one "x" representing the full cross-classification defining household types. We observe n_{ijr} if $r = 1$ or if $r = 0$ and $i \in S$, but only n_{i+0} if $i \notin S$, where S represents the set of blocks selected for the NRFU sample.

The IPF algorithm to fit this model starts with initial estimates $\hat{m}_{ijr}^0 = 1$ for all i, j, r and contains the following three steps in cycle t :

$$\text{Step 1: } \hat{m}_{ijr}^{t+\frac{1}{3}} = \begin{cases} \hat{m}_{ijr}^t \left(\frac{n_{i+r}}{\hat{m}_{i+r}^t} \right) & \text{if } i \in S \text{ or if } i \notin S, r = 1 \\ \hat{m}_{ijr}^t & \text{if } i \notin S, r = 0 \end{cases}$$

$$\text{Step 2: } \hat{m}_{ijr}^{t+\frac{2}{3}} = \begin{cases} \hat{m}_{ijr}^{t+\frac{1}{3}} \left(\frac{n_{ij+}}{\hat{m}_{ij+}^{t+\frac{1}{3}}} \right) & \text{if } i \in S \\ \hat{m}_{ijr}^{t+\frac{1}{3}} \left(\frac{n_{ij1}}{\hat{m}_{ij1}^{t+\frac{1}{3}}} \right) & \text{if } i \notin S \end{cases}$$

$$\text{Step 3: } \hat{m}_{ij1}^{t+1} = \hat{m}_{ij1}^{t+\frac{2}{3}} \left(\frac{n_{+j1}}{\hat{m}_{+j1}^{t+\frac{2}{3}}} \right)$$

$$\hat{m}_{ij0}^{t+1} = \hat{m}_{ij0}^{t+\frac{2}{3}} \left(\frac{\sum_{i \in S} n_{ij0}}{\sum_{i \in S} \hat{m}_{ij0}^{t+\frac{2}{3}}} \right).$$

The scaling factors in each step are based only on observed counts.

These steps are repeated until the estimates of the minimal sufficient statistics for the model, excluding \hat{m}_{i+r} for $i \notin S, r = 0$ (i.e., \hat{m}_{i+r} for $i \in S$ and $i \notin S, r = 1, \hat{m}_{ij+}$ for $i \in S, \hat{m}_{ij1}$ for $i \notin S, \hat{m}_{+j1}$, and $\sum_{i \in S} \hat{m}_{ij0}$) are sufficiently close to their observed values. Denoting the step at which this occurs as t^* , the final step in this algorithm is to set

$$\hat{m}_{ijr}^{t^*+1} = \begin{cases} \hat{m}_{ijr}^{t^*} \left(\frac{n_{i+r}}{\hat{m}_{i+r}^{t^*}} \right) & \text{if } i \notin S, r = 0 \\ \hat{m}_{ijr}^{t^*} & \text{otherwise,} \end{cases}$$

to ensure that estimated block \times response margin ($i * r$) for $i \notin S, r = 0$ equals the observed margin.

This IPF algorithm produces estimates that converge to the maximum likelihood estimates of the model parameters (Zanutto 1998, Part 1, Appendix A). The second case in Step 2 is not needed to maximize the likelihood but is included to obtain predictions for the nonsample nonrespondent cells (i.e., $i \notin S, r = 0$).

References

- Bell, W.R., and Otto, M.C. (1994). Investigation of a model-based approach to estimation under sampling for nonresponse in the decennial census. Unpublished paper presented at the Joint Statistical Meetings, Toronto.
- Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B, Methodological*, 25, 220-233.
- Brackstone, G.J., and Rao, J.N.K. (1976). Raking ratio estimators. *Survey Methodology*, 2, 63-69.
- Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86, 68-78.
- Cox, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.
- Darroch, J.N., and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43, 1470-1480.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-22.
- Fuller, W.A., Isaki, C.T. and Tsay, J.H. (1994). Design and estimation for samples of census nonresponse. In *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: U.S. Bureau of the Census, 289-305.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall Ltd.
- George, J.A., and Penny, R.N. (1987). Initial experience in implementing controlled rounding for confidentiality control. In *Proceedings of the Bureau of the Census Annual Research Conference*, Volume 3. Washington, DC: U.S. Bureau of the Census, 253-262.
- Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Second Edition. New York: John Wiley & Sons, Inc.
- Meng, X.-L., and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267-278.
- Oh, H.L., and Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys* (Eds. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press, 143-184.
- Purcell, N.J., and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, 48, 3-18.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Redfern, P. (1989). European experience of using administrative data for censuses of population: The policy issues that must be addressed. *Survey Methodology*, 15, 83-99.
- Rubin, D.B., and Schenker, N. (1987). Interval estimation from multiply-imputed data: A case study using census agriculture industry codes. *Journal of Official Statistics*, 3, 375-387.
- Schafer, J.L. (1995). Model-based imputation of census short-form items. In *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: Bureau of the Census, 267-299.
- Schindler, E. (1993). Sampling for the count; sampling for non-mail returns. Unpublished report, U.S. Bureau of the Census.
- U.S. Bureau of the Census (1997a). Census 2000 operational plan. Washington, DC.
- U.S. Bureau of the Census (1997b). Report to Congress—the plan for Census 2000. Washington, DC.
- Wilkinson, G.N. and Rogers, C.E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, 22, 392-399.
- Zanutto, E. (1998). *Imputation for Unit Nonresponse: Modeling Sampled Nonresponse Follow-up, Administrative Records, and Matched Substitutes*. Ph.D. thesis, Harvard University, Cambridge, Massachusetts.
- Zanutto, E., and Zaslavsky, A.M. (1995a). A model for imputing nonsample households with sampled nonresponse follow-up. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 608-613.
- Zanutto, E., and Zaslavsky, A. M. (1995b). Models for imputing nonsample households with sampled nonresponse follow-up. In *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: U.S. Bureau of the Census, 673-686.

- Zanutto, E., and Zaslavsky, A.M. (2002). Using administrative records to improve small area estimation: An example from the U.S. Decennial Census. *Journal of Official Statistics*, 18, 559-576.
- Zaslavsky, A.M. (1988). Representing local area adjustments by reweighting of households. *Survey Methodology*, 14, 265-288.
- Zaslavsky, A.M. (1993). Combining census, dual-system, and evaluation study data to estimate population shares. *Journal of the American Statistical Association*, 88, 1092-1105.
- Zaslavsky, A.M. (2004). Representing the Census undercount by multiple imputation of households. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (Eds. A. Gelman and X.-L. Meng). West Sussex, England: John Wiley & Sons, Inc. 129-140.
- Zhang, L.-C., and Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society, Series B*, 66, 479-496.

The 2006 Reverse Record Check Sample Allocation

Alain Th  berge ¹

Abstract

Sample allocation can be optimized with respect to various goals. When there is more than one goal, a compromise allocation must be chosen. In the past, the Reverse Record Check achieved that compromise by having a certain fraction of the sample optimally allocated for each goal (for example, two thirds of the sample is allocated to produce good-quality provincial estimates, and one third to produce a good-quality national estimate). This paper suggests a method that involves selecting the maximum of two or more optimal allocations. By analyzing the impact that the precision of population estimates has on the federal government's equalization payments to the provinces, we can set four goals for the Reverse Record Check's provincial sample allocation. The Reverse Record Check's subprovincial sample allocation requires the smoothing of stratum-level parameters. This paper shows how calibration can be used to achieve this smoothing. The calibration problem and its solution do not assume that the calibration constraints have a solution. This avoids convergence problems inherent in related methods such as the raking ratio.

Key Words: Calibration; Raking ratio; Reverse record check; Sample allocation; Smoothing.

1. Introduction

The Canadian Census of Population is conducted every five years, most recently in 2001. The Reverse Record Check (RRC) measures the undercoverage and part of the overcoverage in the Census. For the next RRC in 2006, it is hoped that most of the census overcoverage will be measured by another survey, the Automated Match Study, which is more efficient for this task. This should make it possible to optimize the RRC sample allocation for undercoverage measurement. RRC coverage estimates are used in conjunction with census counts to produce population estimates. The population estimates are used for various purposes; for example, the federal Department of Finance uses them to calculate the equalization payments that the federal government makes to the provincial governments.

Traditionally, one consideration in allocating the RRC sample among the provinces has been to balance the need for a good-quality estimate of the national rate of persons missed by the Census and the need for good-quality estimates of provincial rates for use in producing Statistics Canada's population estimates.

It was hoped that this approach would also meet the need for good-quality equalization payment estimates (they are estimates because they depend on population estimates), but this has never been verified. The federal government makes equalization payments to the have-not provinces. In this paper, we examine the impact that the provincial sample allocation has on the quality of equalization payment estimates.

If the variance of a variable of interest is the same in every province, we can obtain an optimal allocation for a

minimum-variance national estimate if the sample size is proportional to the frame size for province p , N_p . An allocation that produces provincial estimates of equal variance is one where the sample size is constant (proportional to N_p^0). One way often used to balance the two needs is to make the sample size proportional to $N_p^{1/2}$. A different method of achieving this balance has been used in the past by the RRC: part of the sample is allocated so as to yield provincial estimates of equal variance, and the other part is allocated so as to produce a minimum-variance national estimate. Traditionally, about two thirds of the sample is allocated in such a way as to produce provincial estimates of equal variance.

In this paper, we propose a new method of obtaining a provincial allocation which balances two or more goals. That method involves computing a distinct allocation for each goal, possibly with a different total sample size for each allocation; we obtain the final allocation, which should satisfy every goal, by taking the maximum sample size over each of the distinct allocations for each province.

The optimal subprovincial allocation is simply given by the Neyman allocation. The difficulty lies in predicting the variance in relatively small strata, or more precisely, in predicting the totals (number of persons missed by the Census, number of RRC non-respondents) on which the variance depends. For each province, the approach taken in this paper is to start with more stable national values at the cell level (age \times sex \times marital status) and scale them so that the totals agree with the provincial values for each age group, for each sex and for each marital status. This goal is reminiscent of an iterative raking procedure introduced by Deming and Stephan (1940), also used by Brackstone and Rao (1976). Deville and S  r  ndal (1992) showed how

1. Alain Th  berge, Social Survey Methods Division, Statistics Canada, R.H. Coats Bldg, 15th Floor, Ottawa, Ontario, Canada, K1A 0T6.

calibration can be used to achieve the same result. In the case of the RRC, calibration will be used even though the cells cannot be put in a convenient three-dimensional matrix because the age groups differ for each marital status. The raking ratio method sometimes fails to converge because the constraints cannot be satisfied. By stating the calibration problem as in Théberge (1999), we allow for the possibility that the constraints are inconsistent, and this does not cause convergence problems. In addition, if we use the Moore-Penrose inverse as part of the solution, the constraints can be linearly dependent.

In the next section, we will explore the relationship between population estimates and equalization payments. As we will see, the sample allocation problem entails balancing four goals. In Section 3, we use an approximate variance formula that relies on a design effect to determine the optimal allocation for each goal. We determine the value of the design effect empirically in Section 4. Section 5 explains how a final allocation can balance individual allocations for separate goals. Finally, the subprovincial allocation is addressed in Section 6. The sample allocation for the three territories is not discussed in this paper.

2. Impact of Population Estimates on Equalization Payments

Statistics Canada is responsible for producing population estimates. One important use of those population estimates is in computing the equalization payments made by the federal Department of Finance. Although Statistics Canada is not directly concerned with the formula for equalization payments, it is still relevant to examine how the precision of the population estimates affects the precision of the equalization payments. The impact that the sample allocation has on the precision of the population estimates has been studied for many years; in this paper, we will also examine how the sample allocation affects the precision of the equalization payments.

The RRC is the survey used to measure the rate of persons missed by the Census. Traditionally, the RRC's sample allocation has been designed to achieve a compromise between having a minimum variance for the national estimated undercoverage rate (goal I) and having equally low variances for the estimated undercoverage rates of the provinces (goal II). Two more goals will be added as we examine the impact that the sample allocation has on the precision of the equalization payments.

The formula used to calculate the equalization payments, before any smoothing based on moving average, is

$$E_p = \sum_{j=1}^{33} \frac{R_{ij}}{T_{ij}} \left(\frac{T_{stdj}}{P_{std}} - \frac{T_{pj}}{P_p} \right) P_p, \quad (2.1)$$

where E_p is the equalization payment for beneficiary province p (at the time of writing, all provinces except Ontario and Alberta), R_{ij} is the total revenue (all provinces) from revenue source j , T_{ij} is the total tax base for revenue source j , T_{stdj} is the tax base of the standard provinces (all provinces except the Atlantic provinces and Alberta) for revenue source j , P_{std} is the population of the standard provinces, T_{pj} is the tax base of beneficiary province p for revenue source j , and P_p is the population of beneficiary province p .

To measure the influence that population estimates have on the equalization payments, we will rewrite equation (2.1) as

$$E_p = \left(\frac{P_p}{P_{std}} \right) C_{std} - K_p, \quad (2.2)$$

where

$$C_{std} = \sum_{j=1}^{33} \frac{R_{ij} T_{stdj}}{T_{ij}}$$

and

$$K_p = \sum_{j=1}^{33} \frac{R_{ij} T_{pj}}{T_{ij}}.$$

We note that the population of Alberta has no impact on the equalization payment of any beneficiary province. The population of Ontario only affects the equalization payment through P_{std} . For the Atlantic provinces, their equalization payment varies linearly with their population, since their population does not affect P_{std} . If we assume P_{std} is known, we can say that an error of one person in a beneficiary province's population has an impact of C_{std} / P_{std} dollars on its equalization payment, for any beneficiary province. This does not mean that the equalization payment of a beneficiary province only depends on its population and not on the population of the standard provinces. However, as we will see, most of the sampling error in the equalization payment comes from the sampling error in the estimate of the beneficiary province's population, and relatively little comes from the sampling error in the estimate of the standard provinces' population.

If symbols with hats represent estimates, then from (2.2),

$$V(\hat{E}_p) \approx C_{std}^2 \frac{1}{P_{std}^2} \left(V(\hat{P}_p) + \left(\frac{P_p}{P_{std}} \right)^2 V(\hat{P}_{std}) - 2 \frac{P_p}{P_{std}} \text{Cov}(\hat{P}_p, \hat{P}_{std}) \right). \quad (2.3)$$

Because stratification is done separately for each province, for a beneficiary province p , which is not one of the standard provinces, we have, ignoring interprovincial migration, $\text{Cov}(\hat{P}_p, \hat{P}_{\text{std}}) = 0$, whereas $\text{Cov}(\hat{P}_p, \hat{P}_{\text{std}}) = V(\hat{P}_p)$ for any of the standard provinces. We can compute an approximation by leaving out the last two terms of (2.3):

$$V(\hat{E}_p) \approx \left(\frac{C_{\text{std}}}{P_{\text{std}}}\right)^2 V(\hat{P}_p). \tag{2.4}$$

Using data from the 2001 RRC, we can verify that the standard deviation of the equalization payment derived from (2.4) differs from that derived from (2.3) by no more than 7%, except for two beneficiary provinces: Newfoundland and Labrador, for which the approximation underestimates the standard deviation by 11%, and Quebec, for which the approximation underestimates the standard deviation by 12%.

As we can see from equation (2.4), a sample allocation that produces equal variances for beneficiary provinces' population estimates also produces equal variances for beneficiary provinces' equalization payment estimates. However, having equal CVs for the beneficiary provinces' population estimates does not guarantee equal CVs for the beneficiary provinces' equalization payment estimates, since from equation (2.2), E_p is not directly proportional to P_p , because K_p is not zero. Having equal CVs for the beneficiary provinces' population estimates is still a goal worth pursuing, since it ensures confidence intervals of equal length for the equalization payment per person. Indeed, because of the use of the approximation (2.4), if the 2001 situation recurs in 2006, the confidence interval for Newfoundland and Labrador will be 11% too short (that is, the precision for the equalization payment per person will be poorer than for other beneficiary provinces), while the confidence interval for Quebec will be 12% too long (that is, the precision for the equalization payment per person will be greater than for other beneficiary provinces). Also, if we ignore interprovincial migration, then the provincial population estimates are independent and the variance of the total equalization payment is minimized if and only if the variance of the total population of beneficiary provinces is minimized.

We are attempting to find a provincial sample allocation that minimizes the variance of the total equalization payment or, equivalently, the variance of the total population of beneficiary provinces (goal III). We also want to find a provincial sample allocation that produces equal CVs for each beneficiary province's population estimate (goal IV), in order to achieve equally good precision for the equalization payment per person.

Most of the variance in population estimates is due to the variance in the undercoverage estimates. If we ignore the contribution that overcoverage makes to the variance of the

population estimate, then it is easily verified that the standard error of the estimated undercoverage rate equals the CV of the population estimate. Goals I and II can then be restated as follows: minimize the CV of the national population estimate, and produce provincial population estimates with equal CVs. The difference between goals III and I, and between goals IV and II, is that one applies to beneficiary provinces, and the other to all provinces. In what follows, we will indeed assume that the variance of the population estimates equals the variance of the undercoverage estimates.

The goals of the provincial sample allocation are summarized in Table 2.1.

Table 2.1
The Four Goals of the Provincial Sample Allocation

Goal	Description (equivalent description)
I	Minimize the variance of the estimated national undercoverage rate. (Minimize the CV of the national population estimate.)
II	Produce equal variances for the provinces' estimated undercoverage rates. (Produce provincial population estimates with equal CVs.)
III	Minimize the variance of the total equalization payment (Minimize the variance of the estimated total population of beneficiary provinces).
IV	Produce equal variances for the equalization payment per person for each beneficiary province (Produce equal CVs for the population estimate for each beneficiary province, or produce equal variances for the estimated undercoverage rates of the beneficiary provinces).

3. Optimal Provincial Sample Allocation

In this section, we will first describe the notation we plan to use, and then we will discuss approximate variance formulae for population estimates and estimated undercoverage rates. We will explore the issue of optimality with respect to the four goals mentioned above.

Five sample frames are used for the RRC in the provinces: the census frame (people enumerated in the previous census), the birth frame (intercensal births), the immigrant frame (intercensal immigrants), the non-permanent resident frame and the "missed" frame. The "missed" frame is made up of the sampled persons of the previous RRC who were missed by the previous census. With their weights, they represent the subpopulation of enumerable persons not covered by any of the other four frames. Each frame within each province is stratified separately. A stratified random sample is selected in each frame. All persons from the "missed" frame are included in the sample.

Let U_{hp} be the number of undercovered persons in stratum h who are classified in province (of classification) p . Similarly, let E_{hp} and O_{hp} be, respectively, the number of enumerated and overcovered persons in stratum h who are classified in province p , and $P_{hp} = U_{hp} + E_{hp} - O_{hp}$. The undercoverage rate for province p can then be written as

$$R_{.p} = U_{.p} / P_{.p}, \quad (3.1)$$

where $U_{.p} = \sum_h U_{hp}$ and $P_{.p} = \sum_h P_{hp}$. We see that $P_{.p}$ equals P_p as defined in the preceding section.

An estimator of the undercoverage rate for province p is

$$\hat{R}_{.p} = \hat{U}_{.p} / \hat{P}_{.p}, \quad (3.2)$$

where $\hat{U}_{.p}$ and $\hat{P}_{.p}$ are estimators of $U_{.p}$ and $P_{.p}$ respectively. Linearization gives

$$V(\hat{R}_{.p}) \cong \frac{1}{P_{.p}^2} \left[V(\hat{U}_{.p}) + \frac{U_{.p}^2}{P_{.p}^2} V(\hat{P}_{.p}) \right]. \quad (3.3)$$

The second term in brackets is negligible in comparison to the first; therefore,

$$V(\hat{R}_{.p}) \cong \frac{1}{P_{.p}^2} \sum_h \frac{U_{hp}(N_h - U_{hp})}{n_h}, \quad (3.4)$$

where N_h is the size of stratum h , and n_h is the sample size in stratum h . This ignores the finite population correction factor. In what follows, we will assume that there is no non-response and that there is only one stratum per province of selection (no stratification by frame, age, sex, etc.). This assumption will of course be dropped in Section 6, which deals with sample allocation to subprovincial strata. To compensate for the effects of subprovincial stratification and non-response, we introduce a design effect, D_h . We assume that this design effect varies only with stratum h ; in particular, the same design effect is used to represent the variance of the estimated number of persons selected in stratum h who are undercovered in province p , for all p . The variance (3.4) can be approximated by

$$V(\hat{R}_{.p}) \cong \frac{1}{P_{.p}^2} \sum_{h=1}^{10} \frac{D_h U_{hp}(N_h - U_{hp})}{n_h}, \quad (3.5)$$

and

$$V(\hat{U}_{.p}) \cong \sum_{h=1}^{10} \frac{D_h U_{hp}(N_h - U_{hp})}{n_h}, \quad (3.6)$$

where the summation this time is over the provinces of selection.

Goal I:

From (3.5), we have

$$V(\hat{R}_{.p}) \cong \frac{1}{P_{.p}^2} \sum_{h=1}^{10} \frac{D_h U_{hp}(N_h - U_{hp})}{n_h}, \quad (3.7)$$

where $P_{.p} = \sum_{h=1}^{10} P_{hp}$, $\hat{R}_{.p} = \hat{U}_{.p} / \hat{P}_{.p}$, $U_{.p} = \sum_{h=1}^{10} U_{hp}$ and $U_{hp} = \sum_{p=1}^{10} U_{hp}$. This variance of the national estimated undercoverage rate will be minimized if n_h is proportional to $\sqrt{D_h U_{hp}(N_h - U_{hp})} = N_h \sqrt{D_h R_{hp}(1 - R_{hp})}$, where $R_{hp} = U_{hp} / N_h$. Therefore, the optimal allocation for goal I of a sample of total size n_I is

$$n_{pI} = n_I \left[\frac{N_p \sqrt{D_p R_{.p}(1 - R_{.p})}}{\sum_{p=1}^{10} N_p \sqrt{D_p R_{.p}(1 - R_{.p})}} \right] \quad p = 1, \dots, 10. \quad (3.8)$$

This is an improvement over the formula used for the 2001 RRC (see Clark 2000), where no design effect was applied to the part of the sample allocated to provide the best Canada-level estimate. In addition, for the 2001 RRC, n_p was proportional to the projected population in province p . It makes sense for n_p to depend on the size of the provincial frames; it should also depend on the provincial distribution of the undercoverage.

Goal II:

We can use equation (3.5) to compute the values of n_h that yield the same variance for the estimated provincial undercoverage rates. That problem has 10 equations in 10 unknowns. There is also another difficulty: obtaining sufficiently precise estimates of the U_{hp} for $p \neq h$, especially if p is a small province. Although in many cases it is reasonable to assume that the rate of undercovered persons in a small province p , $R_{.p} = U_{.p} / P_{.p}$, that was observed in one census, is a good predictor of the rate in the next census, the individual values of the U_{hp} for $p \neq h$ are harder to estimate and still harder to predict. Instead, we will assume that $U_{hp} = 0$ for $p \neq h$ and that $U_{pp} = U_{.p}$, which will mitigate the effect that outliers have on the expected variances. The provincial estimates of the undercoverage rate will then be of equal variance, if n_h , for $h=p$, is proportional to $(1/P_{.p}^2) D_p U_{.p}(N_p - U_{.p}) = D_p R_{.p}(N_p / P_{.p} - R_{.p})$. Therefore, the optimal allocation for goal II of a sample of total size n_{II} is

$$n_{pII} = n_{II} \left[\frac{D_p R_{.p}(N_p / P_{.p} - R_{.p})}{\sum_{p=1}^{10} D_p R_{.p}(N_p / P_{.p} - R_{.p})} \right] \quad p = 1, \dots, 10. \quad (3.9)$$

Note that in the 2001 RRC, for the part of the sample allocated to ensure equal precision of the provincial estimates, the sample sizes were set proportional to $D_p \hat{R}_{.p}(1 - \hat{R}_{.p})$ (see Clark 2000). Using $N_p / P_{.p}$ instead of

1 takes into account not only those units which are in the province's frame and leave the province's population but also those units of the province's population that are not in the province's frame, leaving the design effect to account only for non-response and the sample design. In 2001, adjustment for frame units leaving the population was made through the design effect, and no adjustment was made for population units not in the frame.

Goal III:

The estimate of the total population of beneficiary provinces has a variance equal to

$$V(\hat{P}_{ben}) = V(\hat{U}_{ben}) \cong \sum_{h=1}^{10} \frac{D_h U_{hben} (N_h - U_{hben})}{n_h}, \quad (3.10)$$

where $P_{ben} = \sum_{p=1}^8 P_p$, $U_{hben} = \sum_{p=1}^8 U_{hp}$ and $U_{ben} = \sum_{p=1}^8 U_p$ are sums over the eight beneficiary provinces (we assume that the beneficiary provinces are numbered $p = 1, \dots, 8$, and the non-beneficiary provinces are numbered $p = 9, 10$). Equation (3.10) is minimized if n_h , for $h = 1, \dots, 10$, is proportional to $\sqrt{D_h U_{hben} (N_h - U_{hben})} = N_h \sqrt{D_h R_{hben} (1 - R_{hben})}$, where $R_{hben} = U_{hben} / N_h$. Therefore, the optimal allocation for goal III of a sample of total size n_{III} is

$$n_{pIII} = n_{III} \frac{N_p \sqrt{D_p R_{pben} (1 - R_{pben})}}{\sum_{p=1}^{10} N_p \sqrt{D_p R_{pben} (1 - R_{pben})}} \quad p = 1, \dots, 10. \quad (3.11)$$

Note that because units selected in one province can be classified in another province, R_{pben} , and n_{pIII} , are not necessarily zero when p is a non-beneficiary province.

Goal IV:

From equation (3.6), we have

$$CV(\hat{P}_p) \cong \frac{1}{P_p} \sqrt{\sum_{h=1}^{10} \frac{D_h U_{hp} (N_h - U_{hp})}{n_h}}. \quad (3.12)$$

We can use this equation to compute the values of n_h that yield the same coefficient of variation for the beneficiary provinces' population estimates. That problem has eight equations in eight unknowns. Again here, we have a second difficulty: obtaining sufficiently precise estimates of the U_{hp} for $p \neq h$, especially if p is a small province. As we did for goal II, we will assume instead that $U_{hp} = 0$ for $p \neq h$ and that $U_{pp} = U_p$. Beneficiary provinces' population estimates will then have equal coefficients of

variation if n_h , for $h = p$, is proportional to $(1/P_p^2) D_p U_p (N_p - U_p) = D_p R_p (N_p / P_p - R_p)$. Therefore, the optimal allocation for goal IV of a sample of total size n_{IV} is

$$n_{pIV} = n_{IV} \frac{D_p R_p (N_p / P_p - R_p)}{\sum_{p=1}^8 D_p R_p (N_p / P_p - R_p)} \quad p = 1, \dots, 8 \quad (3.13)$$

with the two non-beneficiary provinces having $n_{pIV} = 0$, $p = 9, 10$.

It is worth noting that n_{pII} / n_{pIV} is constant for all eight beneficiary provinces. This shows that goal II (equal precision of the estimated provincial undercoverage rates), which is a traditional goal of the RRC sample allocation, largely overlaps with goal IV (equal precision of the beneficiary provinces' equalization payments per person). We will see in Section 5 that n_{pI} / n_{pIII} , for the eight beneficiary provinces, is nearly constant as well. This shows that goal I (maximum precision of the estimated national undercoverage rate), which is a traditional goal of the RRC sample allocation, largely overlaps with goal III (maximum precision of the total equalization payments).

4. Design Effect

Standard errors for the 2001 RRC estimates were computed using the Generalized Estimation System. Those standard errors take into account the RRC's sampling plan and non-response by assuming that the respondents are selected with a multi-stage sampling plan. A comparison of the standard error derived from (3.6) and the standard error computed by the Generalized Estimation System is presented in Table 4.1. A design effect equal to the inverse of the cube of the response rate for the province of selection was used for this comparison.

The table shows that the standard error for Prince Edward Island derived from (3.6) is 39% higher than the standard error computed by the GES; this is due to an outlier which affects the equation (3.6) estimate more than it affects the GES estimate. For most provinces, the equation (3.6) standard error is close to the GES standard error. These empirical results show that the design effect in equations (3.5) and (3.6) is approximately equal to the inverse response rate cubed. This suggests that a sample size of " n " units with response rate " r " yields the equivalent of " $n \times r^3$ " units rather than the expected $n \times r$, because non-respondents are concentrated among persons missed by the Census. The GES takes into account the fact that undercovered persons are less likely to respond. This

decline in precision due to non-response occurs even though the actual sampling plan is more efficiently stratified than the assumed sampling plan of one stratum per province.

Table 4.1
Comparison of Standard Errors

Province	Response rate	D = (response rate) ⁻³	Standard error of under-coverage estimate from (3.6)	Standard error of under-coverage estimate from GES	(3.6) SE / GES SE
N.L.	0.97	1.08	1,783	1,689	1.06
P.E.I.	0.97	1.09	1,021	734	1.39
N.S.	0.95	1.15	3,903	3,955	0.99
N.B.	0.96	1.13	3,272	3,229	1.01
Que.	0.95	1.17	19,915	19,664	1.01
Ont.	0.92	1.28	31,502	31,602	1.00
Man.	0.95	1.15	4,762	5,115	0.93
Sask.	0.96	1.12	3,921	3,840	1.02
Alta.	0.93	1.25	10,493	10,505	1.00
B.C.	0.91	1.34	14,619	14,763	0.99
Can.	0.94	1.20	42,074	42,041	1.00

There have been no similar studies comparing the design effect and the non-response rate in previous RRCs. The weight adjustment method used to compensate for non-response is different, and the nature of non-response is significantly different from what it was before 2001.

5. Final Provincial Sample Allocation and Example

Table 5.1 shows the parameter values that will be used in the example. The values of \hat{N}_p are projections of RRC frame size for 2006; the other parameters are based on 2001 RRC data.

As we might expect, the values of $\hat{R}_{p\text{ben}}$ in Ontario and Alberta show that few units selected in those two provinces are classified as missed by the Census in beneficiary provinces.

The final sample size allocated to province p is simply

$$n_p = \max(n_{pI}, n_{pII}, n_{pIII}, n_{pIV}) \quad p = 1, \dots, 10. \quad (5.1)$$

Whether we use the maximum of the four sizes as in (5.1), a weighted arithmetic mean, or a weighted geometric mean, each method uses four arbitrary parameters (three if the total sample size is fixed). For the maximum method, higher relative values of n_I (or of n_{II} , n_{III} or n_{IV}) make goal I (II, III or IV respectively) more important.

Table 5.2 presents an example with $n_I = 30,000$, $n_{II} = 64,000$, $n_{III} = 25,000$ and $n_{IV} = 48,078$.

The resulting total sample size is 70,028. Figures in bold represent the maximum for the four allocations, n_p . Small changes in n_{III} would affect only the final allocation for Quebec. This suggests that with the sample sizes n_I , n_{II} , n_{III} and n_{IV} as chosen above, the final sample size allocated to Quebec is dictated by goal III (a precise estimate of the total equalization payment). Similarly, the final sample size allocated to Ontario is dictated by goal I (a precise estimate of the national undercoverage rate). The final sample size allocated to Alberta is dictated by goal II (equal variances for the provinces' estimated undercoverage rates). The final sample sizes of the other provinces are dictated both by goal II and by goal IV (equal precision of the estimated equalization payment per person). As noted in Section 3, n_{pII}/n_{pIV} is constant for all eight beneficiary provinces. In the example above, because of the "judicious" choice of n_{IV} , the constant is 1. Lowering n_{IV} would decrease Alberta's final sample size, but not that of other provinces. We note also that n_{pI}/n_{pIII} does not vary much for the eight beneficiary provinces. The addition of goals III and IV (relating to equalization payments) allows us to control Quebec's sample size and Alberta's sample size separately. When only goals I and II were used, Quebec's sample size tended to be closely tied to Ontario's, while Alberta's sample size was closely tied to that of the other provinces.

Table 5.1
Parameter Values

Province	\hat{N}_p	D_p	\hat{P}_p	\hat{R}_p	\hat{R}_p	$\hat{R}_{p\text{ben}}$
N.L.	551,987	1.0804	524,722	0.0339	0.0464	0.0368
P.E.I.	145,173	1.0882	132,473	0.0334	0.0334	0.0307
N.S.	995,651	1.1527	947,099	0.0492	0.0464	0.0440
N.B.	797,488	1.1345	736,129	0.0493	0.0466	0.0440
Que.	8,079,167	1.1740	7,381,352	0.0510	0.0471	0.0460
Ont.	13,423,132	1.2752	11,702,797	0.0653	0.0565	0.0017
Man.	1,262,547	1.1558	1,136,146	0.0466	0.0437	0.0392
Sask.	1,082,238	1.1223	996,562	0.0437	0.0430	0.0402
Alta.	3,373,128	1.2478	3,010,105	0.0490	0.0403	0.0028
B.C.	4,570,444	1.3369	4,014,502	0.0761	0.0669	0.0620
Can.	34,280,955	1.2039	30,581,887	0.0587	0.0524	0.0258

Table 5.2
Provincial Sample Allocation with
 $n_I = 30,000$, $n_{II} = 64,000$, $n_{III} = 25,000$, and $n_{IV} = 48,078$

Province	n_{pI}	n_{pII}	n_{pIII}	n_{pIV}	n_p	n_{pI} / n_{pIII}
N.L.	427	3,816	546	3,816	3,816	0.78
P.E.I.	96	3,956	132	3,956	3,956	0.73
N.S.	796	5,822	1,107	5,822	5,822	0.72
N.B.	634	5,921	881	5,921	5,921	0.72
Que.	6 562	6,399	9,262	6,399	9,262	0.71
Ont.	12,385	9,220	3,148	0	12,385	3.93
Man.	982	5,867	1,331	5,867	5,867	0.74
Sask.	823	5,234	1,139	5,234	5,234	0.72
Alta.	2,622	6,702	1,015	0	6,702	2.58
B.C.	4,673	11,063	6,440	11,063	11,063	0.73
Total	30,000	64,000	25,000	48,078	70,028	

An allocation method that uses equation (5.1) and a table such as Table 5.2 makes it clear why a province's sample has to be a certain size. For example, if we look at the final sample allocation in Table 5.2 and decide that 5,867 observations in Manitoba is insufficient, then we have to specify the goal for which they are insufficient. If we want to improve on the results for goal II (or goal IV), we also have to increase the sample size in all Atlantic provinces and all western provinces (or in all Atlantic provinces and all western provinces except Alberta).

6. Subprovincial Sample Allocation

Although it is evident from equation (3.5) that the subprovincial sample allocation in one province of selection affects the variances of other provinces' estimates, we will try to optimize the allocation in one province only for that province's estimate. In other words, our problem for each province p is to minimize

$$\sum_{h \in \left\{ \begin{array}{l} \text{strata of province} \\ \text{of selection } p \end{array} \right\}} \frac{D_h U_{hp} (N_h - U_{hp})}{n_h} \quad (6.1)$$

subject to the constraint

$$\sum_{h \in \left\{ \begin{array}{l} \text{strata of province} \\ \text{of selection } p \end{array} \right\}} n_h = n_p,$$

where n_p is a previously determined total sample size for province p . Note that the sample size allocated to the "missed" frame is fixed, which means that in what follows, the "missed" frame strata are ignored, and n_p excludes the "missed" frame sample size. The solution to that minimization problem is

$$n_{h^*} = n_p \frac{\sqrt{D_{h^*} U_{h^*p} (N_{h^*} - U_{h^*p})}}{\sum_{h \in \left\{ \begin{array}{l} \text{strata of province} \\ \text{of selection } p \end{array} \right\}} \sqrt{D_h U_{hp} (N_h - U_{hp})}} \quad (6.2)$$

for each stratum h^* in province of selection p .

As we saw in Section 4, there is empirical evidence at the provincial level that the factor D_h is inversely proportional to the cube of the RRC response rate. For the 2001 sample allocation, it was assumed that D_h varied with the inverse of the response rate. To limit the shift of sample, relative to 2001, from strata with a high response rate, such as census frame or birth frame strata, to strata with a low response rate, such as immigrant frame or non-permanent resident frame strata, we will make D_h proportional to the inverse of the square of the response rate in stratum h . Note that here, in contrast to the assumption we made in Section 3, factor D_h compensates only for non-response; it does not compensate for the stratification since it is defined at the stratum level. This is another reason for choosing a factor smaller than the inverse of the cube of the response rate.

As was the case in the 2001 sample allocation, we are faced with the problem of reliably projecting the 2006 values of U_{hp} and D_h for every stratum h . Since the birth frame, the immigrant frame and the non-permanent resident frame each have only one stratum per province, we plan to use the 2006 sizes for those strata and the 2001 undercoverage rates and response rates, along with some ad hoc adjustments for the less populous provinces if necessary. A similar procedure can be used for the Indian reserve strata of the census frame. The other census frame strata are based on sex, marital status (married, not married) and age group. For these strata, using the same age groups for each sex and each marital status, it would be possible, for each province, to rake the *national* projections to margins of *provincial* projections, and use the raked values in equation (6.2). More precisely, to produce projections for U_{hp} for all strata h in province of selection p , we would first take the 2001 estimated rates and the 2006 strata sizes and compute a projection, for each cell (sex \times marital status \times age group), of the number of missed persons, classified in the province where they were selected. Those national figures could populate the cells of a three-dimensional

matrix. Still using the 2001 estimated rates and the 2006 strata sizes, we would then compute a projection for the number of missed persons, classified in province p , in all of the province's strata by sex, then in all of the province's strata by marital status, and finally in all of the province's strata by age group. Those figures would provide the desired marginal totals of the three-dimensional matrix. Through raking, we could obtain projections for U_{hp} that add up to the desired provincial totals by sex, by marital status and by age group. We can avoid convergence problems, simplify programming and enhance flexibility if we replace raking; we can do so by solving a calibration problem. In fact, we need the added flexibility in this case, because the age groups for married persons are not the same as the age groups for not-married persons.

Here is an example of how calibration is used. The method is based on the following result from Théberge (1999).

If we let \mathbf{U} and \mathbf{T} be positive diagonal matrices of dimension n and q respectively, \mathbf{w}_0 a vector of dimension n , \mathbf{A} a $q \times n$ matrix, and \mathbf{b} a vector of dimension q , then among the weight vectors \mathbf{w} of dimension n that minimize $\|\mathbf{Aw} - \mathbf{b}\|_{\mathbf{T}}^2$, the unique weight vector that minimizes $\|\mathbf{w} - \mathbf{w}_0\|_{\mathbf{U}}^2$ is given by

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{U}^{-1} \mathbf{A}' \mathbf{T}^{1/2} (\mathbf{T}^{1/2} \mathbf{A} \mathbf{U}^{-1} \mathbf{A}' \mathbf{T}^{1/2})^{\dagger} \mathbf{T}^{1/2} (\mathbf{b} - \mathbf{Aw}_0), \quad (6.3)$$

where $\|\mathbf{z} - \mathbf{z}_0\|_{\mathbf{F}}^2 = (\mathbf{z} - \mathbf{z}_0)' \mathbf{F} (\mathbf{z} - \mathbf{z}_0)$ is a weighted distance measure between \mathbf{z} and \mathbf{z}_0 , and \mathbf{G}^{\dagger} is the Moore-Penrose inverse of \mathbf{G} .

The equation $\mathbf{Aw} = \mathbf{b}$ forms the set of q calibration constraints. We will set \mathbf{T} equal to the identity matrix in equation (6.3). If the constraints can be satisfied, then the matrix \mathbf{T} is irrelevant; if not, then setting \mathbf{T} equal to the identity matrix has the effect of giving equal importance to each of the q constraints when we minimize the distance between \mathbf{Aw} and \mathbf{b} .

In this case, in projecting the number of missed persons in each stratum of a given province, we have $\mathbf{A} = \mathbf{MX}$, with

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix},$$

$$\mathbf{X} = \text{diag} \begin{pmatrix} x_{FN0-14} \\ x_{FN15-24} \\ x_{FN25-44} \\ x_{FN45+} \\ x_{FM25-34} \\ x_{FM35+} \\ x_{MN0-14} \\ x_{MN15-24} \\ x_{MN25-44} \\ x_{MN45+} \\ x_{MM25-34} \\ x_{MM35+} \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_{FN0-14} \\ w_{FN15-24} \\ w_{FN25-44} \\ w_{FN45+} \\ w_{FM25-34} \\ w_{FM35+} \\ w_{MN0-14} \\ w_{MN15-24} \\ w_{MN25-44} \\ w_{MN45+} \\ w_{MM25-34} \\ w_{MM35+} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_{F..} \\ b_{M..} \\ b_{N..} \\ b_{M..} \\ b_{..0-14} \\ b_{..15-24} \\ b_{..25+} \end{pmatrix},$$

where, for example, $x_{MN25-44}$ is the number of missed persons, classified in the province where they were selected, in the strata of not-married males aged 25 to 44, $w_{MN25-44}$ is the desired weight for that stratum, and $b_{N..}$ is the number of missed persons selected and classified in the province who belong to the "not married" strata. All persons aged 0 to 24 are in "not married" strata regardless of their actual marital status. Note that in calculating both the national figures, \mathbf{X} , and the provincial figures, \mathbf{b} , we count only persons who did not move from one province to another, so as to remain consistent with the objective set out at the beginning of this section.

Continuing the parallel with raking, the matrix \mathbf{X} gives the values of the three-dimensional matrix to be raked, except that the elements are arranged in a diagonal matrix; the vector \mathbf{w} provides the final "raking factors" that are applied to the elements of \mathbf{X} to produce the raked values, \mathbf{Xw} ; the constraint is that sums of those raked elements, \mathbf{MXw} , should be as close as possible to the desired "margins" given by the vector \mathbf{b} ; and \mathbf{w} should be as close as possible to \mathbf{w}_0 described below.

By choosing the vector \mathbf{w}_0 so that every element is equal to a constant factor, we can scale the national figures down to figures that are more appropriate for the province. We can do this if we want the weighted national figures to add up to the provincial marginal totals, with weights that are as close as possible to a constant, in order to preserve the more reliable national distribution. The national distribution of missed persons may not be appropriate if the distribution of strata sizes is not the same for Canada as it is for the province. Therefore, a better alternative is to set the \mathbf{w}_0 element that corresponds to stratum h^* to

$$w_{0h^*} = N_{h^*} / \sum_{h \in S_{h^*}} N_h, \quad (6.4)$$

where S_{h^*} is the set of the 10 strata (one per province) similar to stratum h^* (for example, the 10 strata of not-married males aged 15 to 24).

We could remove two constraints because the corresponding rows of \mathbf{M} are linear combinations of the others (for example, the fourth row and the last row), but the solution (6.3) is sufficiently general that their removal is unnecessary. With $\mathbf{A} = \mathbf{MX}$, $\mathbf{U} = \mathbf{X}$ and \mathbf{T} equal to the identity matrix, (6.3) simplifies to

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{M}'(\mathbf{MXM}')^{\dagger}(\mathbf{b} - \mathbf{MXw}_0). \quad (6.5)$$

The smoothed values for each stratum are the elements of the vector \mathbf{Xw} .

A similar problem can arise for non-respondents when we want to smooth the sample design's effects.

7. Conclusion

There is much overlap between the two traditional goals of RRC sample allocation, which are to obtain a minimum variance for the national estimated undercoverage rate (goal I) and to obtain equal variances for the estimated provincial undercoverage rates (goal II), and the two additional goals considered in this paper, which are to minimize the variance of the total equalization payment (goal III) and to obtain equal CVs for the beneficiary provinces' population estimates (goal IV). Nevertheless, the explicit consideration of those two additional goals may allow the sample sizes for Quebec and Alberta to vary independently from those of the other provinces. The method suggested in this paper to achieve a compromise between different allocations that is optimal with respect to the various goals, is to take, for each province, the maximum sample size over each of the distinct allocations. The method provides a more direct justification for the allocation.

A comparison of the GES standard errors with the standard errors derived from the approximation formula (3.6) shows for the 2001 RRC, n sampled units with a response rate of r are equivalent to only $n \times r^3$ full-response units.

Optimal subprovincial allocation requires smoothing of provincial parameters at the age \times sex \times marital status level. Calibration can be a convenient method to scale more stable national age \times sex \times marital status values so that they add up to provincial age values, sex values and marital status values. The method's principal goal is reminiscent of the principal goal of the raking ratio method, but a solution such as the one described in Th  berge (1999), which deals with the possibility that the constraints may not be satisfied, avoids convergence problems. In addition, using the Moore-Penrose inverse prevents collinearity problems.

References

- Brackstone, G.J., and Rao, J.N.K. (1976). Raking ratio estimators. *Survey Methodology*, 2, 63-69.
- Clark, C. (september 2000). 2001 Reverse Record Check: Provincial and Territorial sample allocation. Document non publi  . Ottawa. Statistique Canada.
- Deming, W.E., and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11, 427, 444.
- Deville, J.-C., and S  r  dal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Th  berge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

Sample Size Calculation for Small-Area Estimation

Nicholas Tibor Longford¹

Abstract

We describe a general approach to setting the sampling design in surveys that are planned for making inferences about small areas (sub-domains). The approach requires a specification of the inferential priorities for the areas. Sample size allocation schemes are derived first for the direct estimator and then for composite and empirical Bayes estimators. The methods are illustrated on an example of planning a survey of the population of Switzerland and estimating the mean or proportion of a variable for each of its 26 cantons.

Key Words: Efficiency; Inferential priority; Sample size allocation; Small-area estimation.

1. Introduction

Sampling design is a key device for efficient estimation and other forms of inference about a large population when the resources available do not permit collecting the relevant information from every member of the population. In this context, efficiency is interpreted as the optimal combination of a sampling design and an estimator of a population quantity θ . By optimum we understand minimum mean squared error, although the development presented in this paper can be adapted for other criteria. The pool of the possible sampling designs is delimited by the resources, and these are usually expressed in terms of a fixed sample size. This is not always appropriate because the designs may not entail identical average costs per subject. However, within a limited range of designs, this issue can be ignored.

The problem of setting the sampling design for the purpose of efficient estimation of a single quantity is well understood, and solutions are available for many commonly encountered settings. Most of them involve a univariate constrained optimisation problem. Setting the sampling design for estimating several quantities represents a quantum leap in complexity, because the problem involves several factors, typically one for each quantity. It is essential to optimise the design simultaneously for all the factors, because the goals of efficient inference about the target quantities may be in conflict. For example, in small-area estimation, a more generous allocation of the sample size to one area has to be compensated by a less generous allocation to one or several other areas.

Small-area statistics have become an important research topic in survey methods in the last few decades (Fay and Herriot 1979; Platek, Rao, Särndal and Singh 1987; Ghosh and Rao (1994), Longford 1999; and Rao 2003), stimulated by increasing interest of government agencies, the advertising and marketing industry and the financial and insurance sector. At present, many large-scale surveys are

designed for estimating national quantities but, sometimes almost as an afterthought, are used for inferences about small areas. This would be appropriate if the sampling designs optimal for small-area and national inferences were similar. We illustrate in this paper that this is not the case and that sampling design can be effectively targeted for small-area estimation, taking into account the goal of efficient estimation of national quantities. To avoid the trivial case, we assume that the areas have unequal population sizes. We apply the methods to the problem of planning inferences about the 26 cantons of Switzerland; their population sizes range from 15,000 (Appenzell-Innerrhoden) to 1.23 million (Zürich). The population of Switzerland is 7.26 million.

Literature on the subject of planning surveys for small-area estimation is rather sparse. An important contribution is Singh, Gambino and Mantel (1994). In one of the approaches they discuss, the planned sample size for the Canadian Labor Force Survey is split into two parts. One part is allocated optimally for the purpose of national (domain) estimation and the remainder optimally for small-area estimation. For the latter goal, equal subsample sizes are allocated to each area when the areas have equal within-area variances, the finite population correction can be ignored and the areas have equal survey costs per subject, but also when the targets of inference are area-level means. When the targets are population totals, equal allocation to the areas is not efficient, because it handicaps estimation for more populous areas. Even when proportions or rates (percentages) are estimated, the within-area variances depend on the population proportion, although the dependence is weak when all the proportions are distant from zero and unity. For more recent developments in sampling design for small-area estimation, see Marker (2001).

The next section describes the proposed approach based on minimising the weighted sum of the sampling variances (mean squared errors) of the planned estimators, with the

1. Nicholas Tibor Longford, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramón Trias Fargas 25-27, 08005 Barcelona, Spain. E-mail: NTL@SNTL.co.uk.

weights specified to reflect the inferential priorities. It is applied first to direct estimation of the area-level quantities. Then it is extended to incorporate the goal of national estimation, and, finally, to composite estimation in section 3. The concluding section 4 contains a discussion.

The remainder of this section introduces the notation used in the rest of the paper. We assume that area-level population quantities θ_d , $d=1, \dots, D$, are estimated by $\hat{\theta}_d$ with respective mean squared errors (MSE) v_d that are functions of the within-area subsample sizes n_d ; $v_d = v_d(n_d)$. The overall sample size is denoted by n , and is assumed to be fixed. The population sizes are denoted by N (overall) and N_d (for area d). For brevity, we denote $\mathbf{n} = (n_1, \dots, n_D)^T$. Most population quantities θ are functions of a single variable, such as its mean, total, and the like. The variable may be recorded in the survey directly, or constructed from one or several such variables. Although our development is not restricted to such quantities, the motivation is more straightforward with them. An estimator of θ_d is said to be *direct* if it is a function of only the variable concerned on subjects in area d .

We assume that each direct estimator considered is unbiased. This is not particularly restrictive, as most direct estimators are naive estimators or are closely related to them. We assume that the sample sizes for the areas are under the control of the survey designer. This is the case in stratified sampling designs in which the strata coincide with the areas. In section 4, we discuss sampling designs in which such control cannot be exercised; they are particularly relevant for divisions of the country into many (hundreds of) areas.

2. Optimal Design for Direct Estimation

We resolve the conflict between the goals of efficient estimation of the area-level quantities θ_d by choosing the area-level sampling design that minimises the weighted sum of the sampling variances (MSEs),

$$\min_{\mathbf{n}} \sum_{d=1}^D P_d v_d, \quad (1)$$

subject to the constraint of fixed overall sample size $n = \mathbf{n}^T \mathbf{1}_D$; $\mathbf{1}_D$ is the vector of unities of length D . The coefficients P_d are called *inferential priorities*. Greater value of P_d (in relation to the values $P_{d'}$, $d' \neq d$) implies a greater urgency to reduce v_d , because the contribution of area d to the sum in (1) is magnified more than for the other areas.

The optimisation problem in (1) is solved by the method of Lagrange multipliers, or simply by substituting $n_1 = n - n_2 - \dots - n_D$, so that the problem then involves $D-1$ functionally unrelated variables. The solution satisfies the condition

$$P_d \frac{\partial v_d}{\partial n_d} = \text{const.}$$

An analytical expression for the optimal subsample sizes n_d cannot be obtained in general, but when $v_d = \sigma_d^2 / n_d$, as in simple random sampling within areas, the solution is proportional to $\sigma_d \sqrt{P_d}$, that is,

$$n_d^* = n \frac{\sigma_d \sqrt{P_d}}{\sigma_1 \sqrt{P_1} + \dots + \sigma_D \sqrt{P_D}}.$$

When the within-area variances σ_d^2 coincide, $\sigma_1^2 = \dots = \sigma_D^2 = \sigma^2$, this simplifies further; the optimal sample sizes are proportional to $\sqrt{P_d}$ and do not depend on σ^2 .

In most contexts, it is difficult to elicit a suitable set of priorities P_d , and so it is more constructive to propose a convenient parametric class of priorities $\mathbf{P} = (P_1, \dots, P_D)^T$ and illustrate their impact on the sample size allocation. We propose the priorities $P_d = N_d^q$ for $0 \leq q \leq 2$. For $q=0$, inference is equally important for every area. With increasing q , relatively greater importance is ascribed to more populous areas. When $v_d = \sigma^2 / n_d$, the optimal sample size allocation for $q=2$, $n_d^* = n N_d / N$, is proportional to the population sizes in the areas, and so the same sampling design is optimal for national and area-level inferences. For $q > 2$ the sample size allocation is even more generous to the most populous areas, at the expense of less populous areas. As this is counterintuitive in the context of small-area estimation, the choice of an exponent $q > 2$ is probably never appropriate. A negative priority exponent q would be suitable for a survey that aims to focus on the least populous areas. Of course, such a design is very inefficient for estimating the national quantity θ , especially when the areas have widely dispersed population sizes.

The inferential priorities P_d may be functions of quantities other than N_d . For example, the sizes of certain subpopulations of focal interest, such as an ethnic minority in the area, may be used instead of N_d , P_d may be defined differently in the country's regions, or the formula for them may be overridden for one or a few areas.

In some publications of survey analyses, an estimate is reported only when it is based on a sufficiently large sample size or its coefficient of variation (the ratio of the estimated standard error and the estimate) is smaller than a specified threshold. If a 'penalty' for not reporting a quantity is specified, it can be incorporated in the definition of the inferential priorities. The difficulty that may arise is that the objective function in (1) is discontinuous and the standard approaches to its optimisation are no longer applicable. The penalty has to be set with care. If it is too low it is ineffective; if it is set too high the solution will prefer reporting estimates for as many areas as possible, but each with sample size or precision that narrowly exceeds the set

threshold. See Marker (2001) for an alternative approach to this problem.

Figure 1 illustrates the impact of the priority exponent q on the sample size allocation for a survey planned in Switzerland, with the aim of estimating the population means of a variable in its 26 cantons, assuming a common within-canton variance σ^2 . The planned overall sample size is $n=10,000$. The curves in either panel connect the optimal sample sizes for each exponent q ; they are drawn on the linear scale (on the left) and on the log scale (on the right). The population sizes are marked on the horizontal bar at the bottom of each plot. On the log scale, the curves are linear. The log scale is useful also because the population sizes of the cantons are more evenly distributed on it.

For $q=0$, each canton is allocated the same sample size, $10,000/26=385$, and for $q=2$ the allocation is proportional to the canton's population size. For intermediate values of q , sample sizes of the least populous cantons are boosted in relation to proportional allocation ($q=2$), at the expense of reduced allocation to the most populous cantons. The subsample sizes depend very little on q for cantons with population of about 250,000, approximately 3% of the national population size.

2.1 The Priority for National Estimation

As the canton-level subsample sizes differ from the proportional allocation for priority exponent $q < 2$, optimal canton-level estimation is accompanied by a loss of

efficiency of the national estimator. Consider the stratified estimator

$$\hat{\theta} = \frac{1}{N} \sum_{d=1}^D N_d \hat{\theta}_d$$

of the national mean θ of a variable, where $\hat{\theta}_d$ are unbiased estimators of the within-canton means of the same variable. Assuming stratified sampling with simple random sampling within strata (cantons), with $\hat{\theta}_d$ set to the within-stratum sample means,

$$\text{var}(\hat{\theta}) = \frac{1}{N^2} \sum_{d=1}^D \frac{N_d^2}{n_d} (1 - f_d) \sigma_d^2,$$

where $f_d = n_d / N_d$ is the finite population correction.

Figure 2 displays the function that relates the standard error $\sqrt{\text{var}(\hat{\theta})}$ to the priority exponent q , calculated assuming $\sigma^2=100$. The standard error is a decreasing function of q ; it decreases more steeply at $q=0$ than at $q=2$, where it is quite flat. For $q=2$, the goals of canton-level and national estimation are in accord, and $\sqrt{\text{var}(\hat{\theta})}=0.100$. For $q=0$, $\sqrt{\text{var}(\hat{\theta})}=0.143$; in this setting, optimality of the small-area estimation exerts a considerable toll on national estimation, equivalent to halving the sample size ($0.143/0.100 \div \sqrt{2}$). For negative q , the toll is even greater.

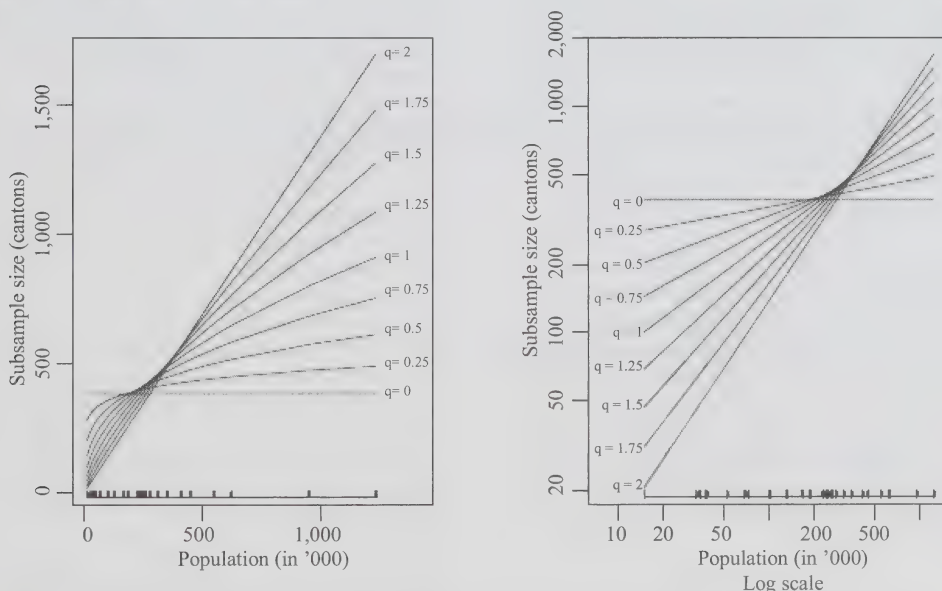


Figure 1. The sample size allocation to the Swiss cantons for a range of priority exponents q . The population sizes of the cantons are marked on the horizontal bar at the bottom of each plot.

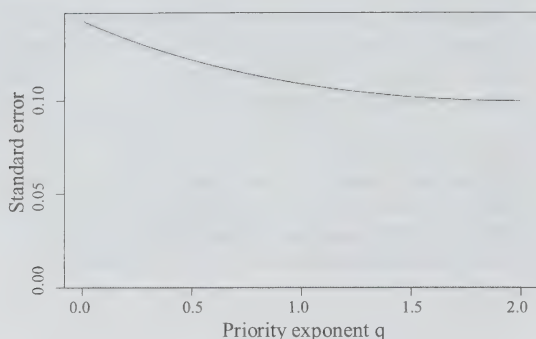


Figure 2. The standard error of the national estimator $\hat{\theta}$ of the mean of a variable, as a function of the exponent q for priorities of the canton-level estimation.

Thus, the need for efficiency of the national estimator can be addressed by increasing the priority exponent. For example, the parties with rival inferential interests may negotiate about how much loss in efficiency of $\hat{\theta}$ can be afforded, and the priority exponent would then be set to match this loss. Alternatively, this loss may be considered by applying the optimal design for area-level estimation. If it is regarded as excessive, q is increased until a balance is struck between the losses of efficiency for national and small-area estimation.

An unsatisfactory feature of these approaches is that they compromise the original purpose of the priorities \mathbf{P} – to reflect the relative importance of the inferences about the distinct small areas. This drawback is addressed by associating $\hat{\theta}$ with a priority, denoted by G , relative to small-area estimation, and considering optimal estimation of the set of D area-level targets θ_d together with the national target θ . Thus, we minimise the objective function

$$\sum_{d=1}^D P_d v_d(n_d) + GP_+ v(\mathbf{n}),$$

where $v = \text{var}(\hat{\theta})$ and $P_+ = \mathbf{P}^T \mathbf{1}_D$. The factor P_+ is introduced to ameliorate the effect of the absolute sizes of P_d and the number of areas on the relative priority G . The priorities P_d can be interpreted only by their relative sizes, as, for any constant $c > 0$, P_d and cP_d correspond to identical sets of priorities for small-area estimation in (1).

When the sampling design within each area is simple random and $\hat{\theta}$ is the standard stratified estimator, the minimum is attained when

$$\sigma_d^2 \frac{P'_d}{n_d^2} = \text{const},$$

where $P'_d = P_d + GP_+ N_d^2 / N^2$. The optimal sample sizes for the areas are

$$n_d^* = n \frac{\sigma_d \sqrt{P'_d}}{\sigma_1 \sqrt{P'_1} + \dots + \sigma_D \sqrt{P'_D}}.$$

This corresponds to an adjustment of the priorities P_d by $GP_+ N_d^2 / N^2$. Note that this adjustment is neither additive nor multiplicative. The priority is boosted more for the more populous areas. As a consequence, the area-level subsample sizes are dispersed more when the relative priority for national estimation is incorporated and the area-level priorities are unchanged. The finite population correction has no impact on n_d^* because it reduces each sampling variance v_d and v by a quantity that does not depend on \mathbf{n} .

The priority G can be set by insisting that the loss of efficiency in estimating the national quantity θ does not exceed a given percentage or that at most a few (or none) of the absolute differences $|P'_d - P_d|$ or log-ratios $|\log(P'_d / P_d)|$ are very large. However, the analytical problem is simple to solve, so the survey management can be presented by the sampling designs that are optimal for a range of values of G .

The dependence of the subsample size on the exponent q and relative priority G is plotted in Figure 3 for the least and most populous cantons, Appenzell-Innerrhoden and Zürich, in the respective panels A and C. Panels B and D plot the same curves as A and C, respectively, on the log scale. Ignoring the goal of national estimation corresponds to $G = 0$ and ignoring the goal of small-area estimation to very large values of G . Throughout, we assume that $n = 10,000$ and $\sigma^2 = 100$, common to all cantons.

For each exponent $q < 2$, the sample-size curve $n_d(G)$ decreases for the less populous and increases for the more populous cantons toward the proportional representation $n_d = nN_d / N$, which corresponds to $q = 2$. On the linear scale, the increase is quite rapid for Zürich for small q and G , whereas the reduction for Appenzell-Innerrhoden is more gradual. As the relative priority G is reduced, the excess sample size is re-distributed from Zürich (and a few other populous cantons) to several less populous cantons.

Figure 4 plots the ‘national’ standard error $\sqrt{\text{var}(\hat{\theta})}$ under the optimal sample allocation for an array of values of q and G . The diagram shows that the standard error of $\hat{\theta}$ is reduced radically by a small increase of G in the vicinity of $G = 0$, whereas for larger values of G it is affected only slightly. For each G , higher priority exponent q is associated with higher precision of $\hat{\theta}$.

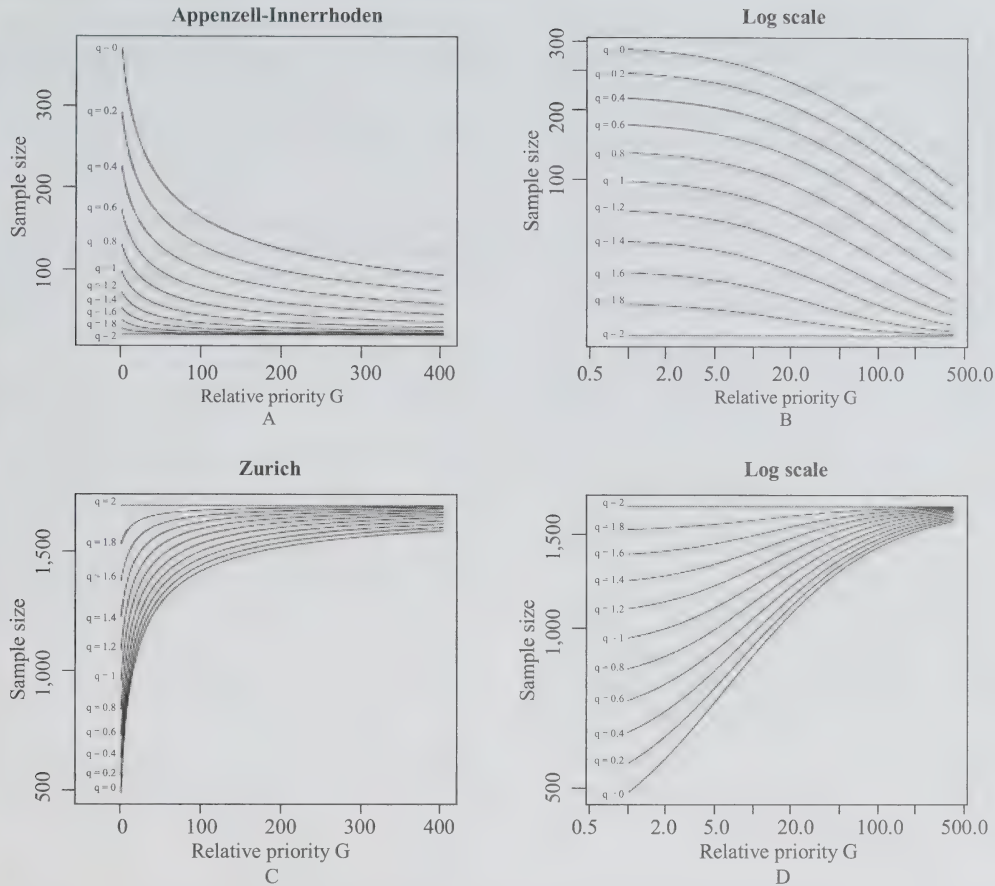


Figure 3. The optimal sample sizes for the direct estimator $\hat{\theta}_d$ for combinations of priority exponents q and relative priorities G for the least and most populous cantons.

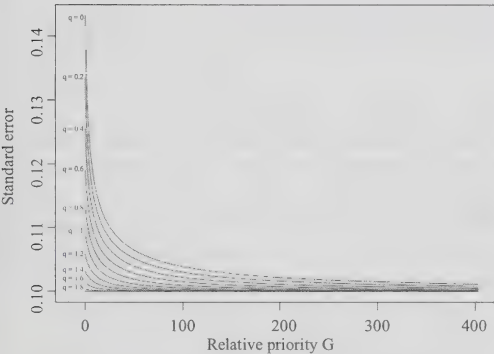


Figure 4. The standard error of the national estimator for the allocation that is optimal under an array of priorities given by q and G .

3. Composite Estimation

The resources available for the conduct of a survey are used most effectively by the optimal combination of a sampling design and estimator(s), and so the sampling design and (the selection of) the estimator should be, in ideal circumstances, optimised simultaneously. This problem is difficult to solve formally in most settings, although some estimators are more efficient than their competitors in a wide range of designs. Composite estimators (Longford 1999, 2004) are one such class. They are convex combinations of the direct small-area and national estimators,

$$\tilde{\theta}_d = (1 - b_d) \hat{\theta}_d + b_d \hat{\theta}, \tag{2}$$

with area-specific coefficients b_d that are estimates of the optimum. The composition $\bar{\theta}_d$ exploits the similarity of the areas; it is particularly effective when the areas have a small between-area variance $\sigma_B^2 = D^{-1} \sum_d (\theta_d - \bar{\theta})^2$, where $\bar{\theta} = D^{-1} \sum_d \theta_d$. This variance is defined over the D population quantities θ_d and is unaffected by the sampling design. In practice, σ_B^2 has to be estimated. When planning a survey, estimates from other surveys of the same or a related population have to be used, and the uncertainty about σ_B^2 addressed. This can be done by sensitivity analysis, exploring the optimal designs for a range of plausible values of σ_B^2 .

If the deviations $\Delta_d = \theta_d - \bar{\theta}$ were known the optimal coefficient b_d in (2) would be, approximately, $b_d^* = \sigma_d^2 / (\sigma_d^2 + n_d \Delta_d^2)$. As Δ_d is not known (otherwise θ_d would be estimated with high precision by $\bar{\theta} + \Delta_d$), we replace Δ_d^2 by its average over the areas, equal to σ_B^2 , yielding the coefficient $b_d = 1 / (1 + n_d \omega_d)$, where $\omega_d = \sigma_B^2 / \sigma_d^2$ is the variance ratio. The variance σ_B^2 also has to be estimated, but when there are many areas it is estimated with precision much higher than most Δ_d^2 are.

If the coefficients b_d are estimated with sufficient precision the composite estimator $\tilde{\theta}_d$ is more efficient than the two constituent estimators $\hat{\theta}_d$ and $\bar{\theta}$. Ignoring the uncertainty about the within- and between-area variances, as well as the national mean $\bar{\theta}$ and the correlation between the national and area-level (direct) estimators, the average MSE of $\tilde{\theta}_d$ is

$$\text{aMSE}(\tilde{\theta}_d) = \frac{\sigma_B^2}{1 + n_d \omega_d}, \quad (3)$$

where ‘aMSE’ denotes the MSE in which Δ_d^2 is replaced by σ_B^2 , its average over the areas. The aMSE in (3) is also an approximation to the conditional variance of the EBLUP estimator of the area-level mean based on the two-level (empirical Bayes) model (Longford 1993, Goldstein 1995, Marker 1999, and Rao 2003). See Ghosh and Rao (1994) for an authoritative review of application of these models to small-area estimation.

For the composite estimators of the area-level means, we search for the sample allocation that minimises the objective function

$$\sum_{d=1}^D P_d \text{aMSE}(\tilde{\theta}_d) + GP_+ v.$$

The solution satisfies the condition

$$\frac{N_d^q \sigma_B^2 \omega_d}{(1 + n_d \omega_d)^2} + GP_+ \frac{N_d^2 \sigma_d^2}{N^2 n_d^2} = \text{const.} \quad (4)$$

This equation does not have a convenient closed-form solution, but iterative schemes can be applied to solve it. The value of n_i determines the remaining sample sizes n_d , and so optimisation corresponds to a one-dimensional search. If the provisional sample sizes \mathbf{n} based on a set value of n_i are too large, $\mathbf{n}^\top \mathbf{1}_D > n$, n_i is reduced and the other sample sizes n_d are calculated by solving (4). Note that the solution depends on the variances σ_d^2 and σ_B^2 . The problem is simplified somewhat when the areas have a common variance $\sigma^2 = \sigma_1^2 = \dots = \sigma_D^2$. Then the solution of (4) depends on the variances only through the ratio $\omega = \sigma_B^2 / \sigma^2$ because σ^2 is a multiplicative factor and has no impact on the optimisation.

By way of an example, suppose $q=1$ and $G=10$ in planning a survey of the population of Switzerland with $n=10,000$, and $\omega=0.10$ is assumed. As the initial solution, we use the allocation optimal for direct estimation with the same values of q and G . One iteration updates the sample size for each canton and, within it, the updating for all but the arbitrarily selected reference canton $d=1$ is also iterative. The reference canton’s provisional subsample size determines the current value of the constant on the right-hand side of (4). Then equation (4) is solved, iteratively, for each canton $d=2, \dots, D$, using the Newton method. In the application, the number of these iterations was in single digits for each canton. Finally, the subsample size for the reference canton is adjusted by the $1/D$ -multiple of the difference between the current total of the subsample sizes and the target total n . The updating of the cantons is itself iterated, but only a few iterations are required to achieve convergence; for example, all the changes in the subsample sizes were smaller than 1.0 after three iterations, and smaller than 0.01 after eight iterations. The convergence is fast because the starting solution is close to the optimum; the largest difference between the two subsample sizes is for Zürich, 20.0 (from 1199.5 at the start to 1219.5 after eight iterations). For Appenzell-Innerrhoden, the sample size is reduced from 81.6 to 73.4. Change by less than unity takes place for five cantons with population sizes in the range 228,000–278,000. Note that the subsample sizes would in practice be rounded, and possibly adjusted further to conform with various survey management constraints.

No priority for national estimation

If national estimation has no priority, $G=0$, equation (4) has the explicit solution

$$n_d^* = \frac{n\omega + D}{\omega} \frac{N_d^{q/2}}{U^{(q)}} - \frac{1}{\omega},$$

where $U^{(q)} = N_1^{q/2} + \dots + N_D^{q/2}$. This allocation is related to the allocation n_d^* , $d=1, \dots, D$, that is optimal for direct estimation of θ_d by the identity

$$n_d^* = n_d^{\dagger} + \frac{1}{\omega} \left(\frac{DN_d^{q/2}}{U^{(q)}} - 1 \right).$$

Hence, when $q > 0$, the allocation optimal for composite estimation is more dispersed than for direct estimation. The break-even population size is $N_T = (U^{(q)} / D)^{2/q}$; areas with population sizes $N_d < N_T$ have smaller subsample sizes for composite than for direct estimation, and areas with greater population sizes have greater subsample sizes. (For $q = 0$, $n_d^* \equiv n / D$). The amount of extra dispersion is inversely proportional to ω .

For $\omega = 0$, the equations for the optimal sampling design lead to a singularity. In this case, each θ_d is estimated efficiently by the national estimator $\hat{\theta}$, and so the design optimal for composite estimation coincides with the design that is optimal for the national estimator ($n_d^* = nN_d / N$). For $q > 0$, the optimal allocation yields negative sample sizes n_d^* when

$$N_d < \left\{ \frac{U^{(q)}}{n\omega + D} \right\}^{2/q}. \quad (5)$$

This (formal) solution is not meaningful. A negative solution should come as no surprise because the aMSE in (3) is an analytical function for $n_d > -1/\omega_d$. For small $\omega > 0$, the aMSE is a shallow decreasing function of the sample size n_d . A negative n_d^* indicates that a (small) canton is not worth sampling because of its low inferential priority P_d . Although additional sample size for a more populous canton d' may yield a smaller reduction of aMSE than it would for a small canton d , its impact is magnified by the larger priority $P_{d'}$.

Positive priority for the national mean

The aMSE in (3) ignores the uncertainty about the national mean θ , and this becomes acute when one of the cantons is not represented in the sample. This deficiency of (3) can be compensated for by setting the relative priority G to a positive value.

Figure 5 summarises the impact of the relative priority G and the priority exponent q on the optimal sample sizes of the least and most populous cantons, together with canton Thurgau which has the 13th (median) largest population size, 228,000. Each setting of q , indicated in the title, and G ,

using different line types, is represented for a canton by a graph of the optimal sample size as a function of the variance ratio ω . The limit of this function for $\omega \rightarrow +\infty$, equal to the sample size optimal for direct estimation, is marked by a bar at the right-hand margin of the panel. For $\omega = 0$, the sampling design optimal for estimation of the national mean θ is obtained. Panels A and B at the top are for the overall sample size $n = 10,000$ and panels C and D for $n = 1,000$.

The diagram shows that the optimal sample sizes are nearly constant in the range $\omega \in (\omega^*, +\infty)$; ω^* increases with q , G and $1/n$. This is a consequence of the relatively large sample size n , which ensures that the subsamples of most cantons are too large for any substantial borrowing of strength across the cantons to take place, unless the cantons are very similar ($\omega < \omega^*$). Most shrinkage coefficients $b_d = 1/(1 + n_d \omega)$ are very small. When $n = 10,000$ is planned, for small values of ω , the optimal sample size increases steeply for the least populous canton and drops precipitously for the most populous canton. Dispersion of the optimal sample sizes increases with q and G , converging to the optimal allocation for estimating the national mean θ , which corresponds to $\omega = 0$. In contrast, the optimal sample sizes are discontinuous at $\omega = 0$ when $G = 0$; the solutions diverge to $-\infty$ for the least populous cantons.

In panels C and D, for $n = 1,000$, the dependence of the sample sizes on ω persists over a wider range of ω because there is a greater scope for borrowing strength across the cantons with the smaller sample sizes. The optimal sample sizes are not monotone functions of ω ; for the least populous cantons there is a dip at small values of ω . The dip is more pronounced for small G and large q , that is, when the disparities of the cantons' priorities are greater and inference about the national mean is relatively unimportant. This phenomenon, somewhat exaggerated by the log-scale of the vertical axis, is similar to the case discussed for $G = 0$. Because of the disparity in the priorities P_d , a small reduction of aMSE for a more populous canton is preferred to a greater reduction for a less populous canton. The dip is present also when $n = 10,000$, but it is so shallow and narrow as to be invisible with the resolution of the graph. Note that the horizontal axes in panels C and D have three times wider range of values of ω than in panels A and B.

In the context of the planned survey, it was agreed that ω is unlikely to be smaller than 0.05. Therefore, the sample size calculations could be based on the direct estimator.

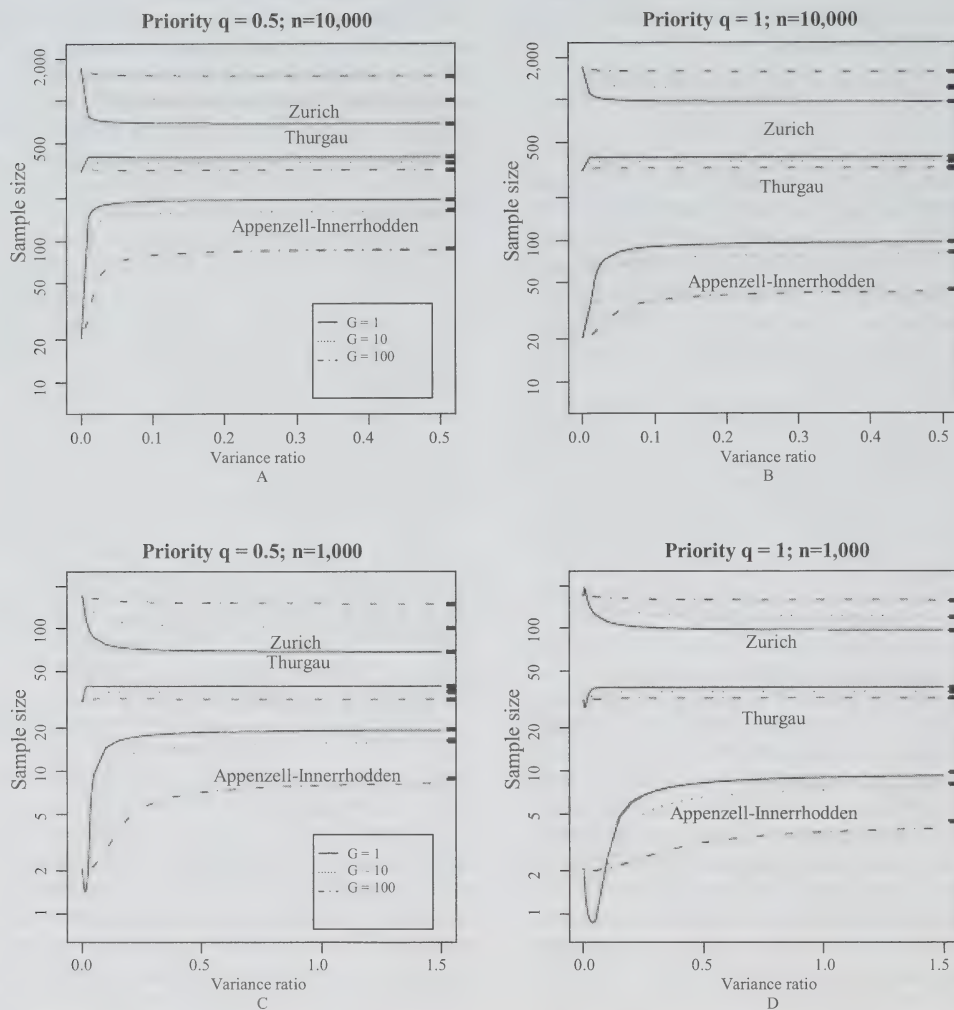


Figure 5. The sample sizes optimal for composite estimation of the population means for three cantons for a range of variance ratios ω , priority exponents $q=0.5$ and $q=1.0$ and relative priorities $G=1, 10$ and 100 . The overall sample sizes are $10,000$ (panels A and B) and $1,000$ (panels C and D).

4. Discussion

The method described in this paper identifies the optimal design for the artificial setting of stratified sampling with simple random sampling within homoscedastic strata. Specifying the priorities for small-area and national estimation is a key element of the method. In practice, the priorities may be difficult to agree on, and some of the assumptions made may be problematic, the assumptions of equal within-stratum variances and simple random sampling

in particular. The method can be extended to more complex estimators, but then the values of further parameters are required. A more constructive approach regards the optimal sampling design for the simplified setting as an approximation to the sampling design that is optimal for the more realistic setting. Even if the optimal sampling design were identified, it could not be implemented literally, because of imperfections in the sampling frame and (possibly) informative and unevenly distributed nonresponse. However, the approach can be applied, in principle, to any small-area

estimator that has an analytical expression for the exact or approximate MSE. This includes all estimators based on empirical Bayes models, to which the composite estimator is closely related. Sampling weights can be incorporated in sample size calculation if they, or their within-area distributions, are known, subject to some approximation, in advance. Sample size calculation for a single (national) quantity entails the same problem.

Although the numerical solution of the problem for composite estimation with a positive priority G is simple and involves no convergence problems, it is advantageous to have an analytical solution, so that a range of scenarios can be explored. The proximity of the solutions for the direct and composite estimation suggests that the allocation optimal for direct estimation may be close to optimum also for composite estimation with realistic values of ω , say, $\omega > 0.05$.

Various management and organisational constraints are another obstacle to the literal implementation of an analytically derived sampling design. In household surveys, it is often preferable to assign an (almost) full quota of addresses to each interviewer, and so sample sizes that are multiples of the quota are preferred. These and numerous other constraints can be incorporated in the optimization problem, although they are often difficult to quantify or the designer may not be aware of them because of imperfect communication. Improvisation, after obtaining the sampling design that is optimal for a simpler setting, may be more practical. Also, priorities, or expert opinion about them, may change over time, even while the survey is being conducted and analysed. Estimates that are associated with standard errors or coefficients of variation greater than a specified threshold are often excluded from analysis reports. Intention to do this can be reflected in sample size calculation by regarding $\hat{\theta}$ as the estimator of θ_d , that is, by setting the associated MSE to the corresponding $\text{aMSE } \sigma_B^2 + \text{var}(\hat{\theta})$ or to another (large) constant.

Although we propose a particular class of priorities for the small areas, no conceptual difficulties arise when another class is used instead. It may depend on several population quantities, not only the population size. In principle, the priorities can also be set for the areas individually, although this is practical only when the number of areas is small. The formula-based and individually set priorities can be combined by adjusting the priorities, such as $P_d = N_d^q$, for a few areas to reflect their exceptional role in the analysis.

Sensitivity analysis, exploring how the sampling design is changed as a result of altered input, is essential for understanding the impact of uncertainty about the estimated parameters (the variance ratio ω in particular) and the arbitrariness, however limited, in how the priorities are set.

For this, an analytically simple solution that can be executed many times, for a range of settings, is preferred to a more complex solution, the properties of which are more difficult to explore.

Multivariate composite estimators exploit the similarity not only across areas, but also across (auxiliary) variables, time, subpopulations, and the like (Longford 1999 and 2005). The aMSEs of these estimators depend on the scaled variance matrix Ω , the multivariate counterpart of ω . Sample size calculation for this method is difficult to implement directly because both variances and covariances in Ω are essential to the efficiency of the estimators. A more constructive approach matches the matrix Ω with a ratio ω that can be interpreted as the similarity of the areas after adjusting for the auxiliary information, as in empirical Bayes methods.

When control over the sample sizes allocated to the areas is not possible sample size calculation is still meaningful as a guide for how the sample sizes should be allocated *on average*. In general, a unit reduction of the sample size is associated with greater loss of precision than a unit increase. Therefore, designs in which the sampling (replication) variance of the subsample sizes $n_d(d \text{ fixed})$ is smaller are better suited for small-area estimation. In designs with large clusters, such variances are large because, at an extreme, an area may not be represented in the survey in some replications and may be over-represented several times in others. Using smaller clusters is in general preferable for small-area estimation if this does not inflate the survey costs and a fixed overall sample size can be maintained.

Acknowledgements

I am grateful to the Deputy Editor and referees for suggesting several improvements but mainly for leading me to discover an error in an earlier version of the manuscript. Discussions with the Polish team in the EURAREA project are acknowledged.

References

- Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.
- Goldstein, H. (1995). *Multilevel Statistical Models*. Second Edition. Edward Arnold, London, UK.
- Longford, N.T. (1993). *Random Coefficient Models*. Oxford University Press, Oxford.

- Longford, N.T. (1999). Multivariate shrinkage estimation of small-area means and proportions. *Journal of the Royal Statistical Society, Series A*, 162, 227-245.
- Longford, N.T. (2004). Missing data and small area estimation in the UK Labour Force Survey. *Journal of the Royal Statistical Society, Series A*, 167, 341-373.
- Longford, N.T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. Springer-Verlag, New York.
- Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.
- Marker, D.A. (2001). Producing small area estimates from national surveys: methods for minimizing use of indirect estimators. *Survey Methodology*, 27, 183-188.
- Platek, R., Rao, J.N.K., Särndal, C.-E. and Singh, M.P. (Eds.) (1987). *Small Area Statistics*. New York: John Wiley & Sons.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-22.

Small Area Estimation Using Area Level Models and Estimated Sampling Variances

Yong You and Beatrice Chapman¹

Abstract

In small area estimation, area level models such as the Fay–Herriot model (Fay and Herriot 1979) are widely used to obtain efficient model-based estimators for small areas. The sampling error variances are customarily assumed to be known in the model. In this paper we consider the situation where the sampling error variances are estimated individually by direct estimators. A full hierarchical Bayes (HB) model is constructed for the direct survey estimators and the sampling error variances estimators. The Gibbs sampling method is employed to obtain the small area HB estimators. The proposed HB approach automatically takes account of the extra uncertainty of estimating the sampling error variances, especially when the area-specific sample sizes are small. We compare the proposed HB model with the Fay–Herriot model through analysis of two survey data sets. Our results have shown that the proposed HB estimators perform quite well compared to the direct estimates. We also discussed the problem of priors on the variance components.

Key Words: Gibbs sampling; Hierarchical Bayes; Prior sensitivity; Sample size; Variance components.

1. Introduction

Sample surveys, for most purposes, are usually designed to provide reliable direct estimates for total populations and large areas by using area-specific sample data. These direct estimates frequently fail to provide reliable estimates for small areas due to very small sample sizes in the areas. Since small area estimates often have unsuitably large standard errors, to gain precision and reliability it is necessary to “borrow strength” from related areas thus increasing the effective sample size to construct indirect estimates for the small areas (Rao 1999). Explicit model-based methods that use supplementary data such as census and administrative data associated with the small areas in explicit models to link the small areas have been widely used in practice to obtain reliable model-based estimators. There are two broad classifications for these models: area level models and unit level models. Area level models are based on area direct survey estimators and unit level models are based on individual observations in the areas. For overviews and appraisals of models for small area estimation, see Rao (1999, 2003). In this paper we study area level models.

To obtain a basic area level model we assume that the small area parameter of interest θ_i is related to area-specific auxiliary data $x_i = (x_{i1}, \dots, x_{ip})'$ through a linear model

$$\theta_i = x_i' \beta + v_i, \quad i = 1, \dots, m, \quad (1)$$

where m is the number of small areas, $\beta = (\beta_1, \dots, \beta_p)'$ is the $p \times 1$ vector of regression coefficients, and the v_i 's are area-specific random effects assumed to be independent and identically distributed (iid) with $E(v_i) = 0$ and $\text{var}(v_i) = \sigma_v^2$. The assumption of normality may also be

included. This model is referred to as a linking model for θ_i .

The basic area level model also assumes that given the area-specific sample size $n_i > 1$, there exists a direct survey estimator y_i (usually design unbiased) for the small area parameter θ_i such that

$$y_i = \theta_i + e_i, \quad i = 1, \dots, m, \quad (2)$$

where the e_i is the sampling error associated with the direct estimator y_i . We also assume that the e_i 's are independent normal random variables with mean $E(e_i | \theta_i) = 0$ and sampling variance $\text{var}(e_i | \theta_i) = \sigma_e^2$. Combining models (1) and (2) lead to a linear mixed area level model

$$y_i = x_i' \beta + v_i + e_i, \quad i = 1, \dots, m. \quad (3)$$

The well-known Fay–Herriot model (Fay and Herriot 1979) in small area estimation has the form of model (3) with the sampling variance σ_e^2 assumed to be known in the model. This is a very strong assumption. Usually a smoothed estimator of σ_e^2 is used in the model and then treated as known. In this paper, we consider the situation where the sampling variances σ_e^2 are unknown and are estimated by unbiased estimators s_i^2 . Following Rivest and Vandal (2002) and Wang and Fuller (2003), we assume that the estimators s_i^2 are independent of the direct survey estimators y_i and s_i^2 has a sampling distribution $d_i s_i^2 \sim \sigma_e^2 \chi_{d_i}^2$, where $d_i = n_i - 1$ and n_i is the sample size for the i^{th} area. For example, suppose we have n_i observations from small area i and these observations are iid $N(\mu_i, \sigma^2)$. Let y_i be the sample mean of the n_i observations. Then $y_i \sim N(\mu_i, \sigma^2/n_i)$ and $\sigma_e^2 = \sigma^2/n_i$. Then we can obtain a direct estimator of σ_e^2 as $s_i^2 = \tau_i^2/n_i$, where τ_i^2 is the sample

variance of the n_i observations. Also y_i and s_i^2 are independent and $(n_i - 1)s_i^2 \sim \sigma_i^2 \chi_{n_i-1}^2$.

We are interested in estimating the small area parameters θ_i . Rivest and Vandal (2002) and Wang and Fuller (2003) obtained the empirical best linear unbiased prediction (EBLUP) estimators of θ_i and the associated mean square error (MSE) approximations assuming that m and n_i are relatively large. In this paper, we consider a hierarchical Bayes (HB) approach using the Gibbs sampling method. An advantage of the HB approach is that it is straightforward, and the inferences for parameters θ_i are “exact” unlike the EBLUP approach. The small area parameter θ_i is estimated by its posterior mean and its precision is measured by its posterior variance. The HB approach automatically takes account of the uncertainties associated with unknown parameters in the model. Section 2 presents the HB area level models and related Gibbs sampling inferences. Section 3 presents two survey data analysis and sensitivity analysis. And finally in section 4, we offer some conclusions and future work directions.

2. Hierarchical Bayes Approach

We now present the area level model (3) and the estimated sampling variances s_i^2 in a HB framework as follows:

Model 1

- $y_i | \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2)$, $i = 1, \dots, m$;
- $d_i s_i^2 | \sigma_i^2 \sim \text{ind } \sigma_i^2 \chi_{d_i}^2$, $d_i = n_i - 1$, $i = 1, \dots, m$;
- $\theta_i | \beta, \sigma_v^2 \sim \text{ind } N(x_i' \beta, \sigma_v^2)$, $i = 1, \dots, m$;
- Priors for the parameters: $\pi(\beta) \propto 1$, $\pi(\sigma_i^2) \sim \text{IG}(a_i, b_i)$, $i = 1, \dots, m$, $\pi(\sigma_v^2) \sim \text{IG}(a_0, b_0)$, where a_i, b_i ($0 \leq i \leq m$) are chosen to be very small known constants to reflect vague knowledge on σ_i^2 and σ_v^2 . IG denotes the inverse gamma distribution.

In Model 1, the sampling variances σ_i^2 are unknown. In practice however, we may have a simpler model by replacing σ_i^2 by its estimate s_i^2 (here s_i^2 is treated as a constant) and obtain the following model:

Model 2

- $y_i | \theta_i \sim \text{ind } N(\theta_i, \sigma_i^2 = s_i^2)$, $i = 1, \dots, m$;
- $\theta_i | \beta, \sigma_v^2 \sim \text{ind } N(x_i' \beta, \sigma_v^2)$, $i = 1, \dots, m$;
- Priors: $\pi(\beta) \propto 1$, $\pi(\sigma_v^2) \sim \text{IG}(a_0, b_0)$.

Model 2 is actually the Fay-Herriot model with sampling variances known as s_i^2 . If area-specific sample sizes n_i are small, using s_i^2 in Model 2 may lead to underestimation of the MSE under the EBLUP approach or the posterior variance under the HB approach. We are interested in

evaluating the effects of using s_i^2 for σ_i^2 in the model. We will obtain the HB estimates of θ_i under both Model 1 and Model 2 and compare the HB estimates through real survey data analysis.

Under the HB approach, we use the posterior mean $E(\theta_i | y)$ as a point estimate for θ_i and the posterior variance $V(\theta_i | y)$ as a measure of variability, where $y = (y_1, \dots, y_m)'$. To estimate $E(\theta_i | y)$ and $V(\theta_i | y)$, we employ the Gibbs sampling method (Gelfand and Smith 1990). From Model 1, we obtain the following full conditional distributions for the Gibbs sampler:

$$\bullet [\theta_i | y, \beta, \sigma_i^2, \sigma_v^2] \sim N(\gamma_i y_i + (1 - \gamma_i) x_i' \beta, \gamma_i \sigma_i^2), \text{ where } \gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_i^2}, i = 1, \dots, m;$$

$$\bullet [\beta | y, \theta, \sigma_i^2, \sigma_v^2] \sim N_p \left(\left(\sum_{i=1}^m x_i x_i' \right)^{-1} \left(\sum_{i=1}^m x_i \theta_i \right), \sigma_v^2 \left(\sum_{i=1}^m x_i x_i' \right)^{-1} \right);$$

$$\bullet [\sigma_i^2 | y, \theta, \beta, \sigma_v^2] \sim \text{IG} \left(a_i + \frac{d_i + 1}{2}, b_i + \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2} \right),$$

where $d_i = n_i - 1$, $i = 1, \dots, m$;

$$\bullet [\sigma_v^2 | y, \theta, \beta, \sigma_i^2] \sim \text{IG} \left(a_0 + \frac{m}{2}, b_0 + \frac{1}{2} \sum_{i=1}^m (\theta_i - x_i' \beta)^2 \right).$$

It is straight forward to draw samples from these full conditional distributions. For implementations, we use $L = 5$ parallel runs each with a “burn-in” length of $B = 1,000$ and Gibbs sampling size of $G = 5,000$. The prior parameters a_i , b_i and a_0, b_0 are chosen to 0.0001. The HB estimator of θ_i under Model 1 is thus obtained as

$$\hat{\theta}_i^{\text{HB}} = (LG)^{-1} \sum_{l=1}^L \sum_{g=1}^G (\gamma_i^{(lg)} y_i + (1 - \gamma_i^{(lg)}) x_i' \beta^{(lg)}), \quad (4)$$

where $\gamma_i^{(lg)} = \sigma_v^{2(lg)} / (\sigma_v^{2(lg)} + \sigma_i^{2(lg)})$, and the posterior variance of θ_i can be estimated by

$$\begin{aligned} \hat{V}(\theta_i) = & (LG)^{-1} \sum_{l=1}^L \sum_{g=1}^G (\gamma_i^{(lg)} \sigma_i^{2(lg)}) \\ & + (LG)^{-1} \sum_{l=1}^L \sum_{g=1}^G (\gamma_i^{(lg)} y_i + (1 - \gamma_i^{(lg)}) x_i' \beta^{(lg)})^2 \\ & - \left\{ (LG)^{-1} \sum_{l=1}^L \sum_{g=1}^G (\gamma_i^{(lg)} y_i + (1 - \gamma_i^{(lg)}) x_i' \beta^{(lg)}) \right\}^2, \end{aligned} \quad (5)$$

where $\{\beta^{(lg)}, \sigma_i^{2(lg)}, \sigma_v^{2(lg)}; g=1, \dots, G; l=1, \dots, L\}$ is the sample generated from the Gibbs sampler. The estimators (4) and (5) are the so-called Rao-Blackwellized HB estimators. The Rao-Blackwellized estimators are more stable in terms of simulation errors as shown, for example, in Gelfand and Smith (1991) and You and Rao (2000).

Now we consider Model 2. The full conditional distributions for the Gibbs sampler under Model 2 are

- $[\theta_i | y, \beta, \sigma_v^2] \sim N(\gamma_i y_i + (1-\gamma_i)x'_i \beta, \gamma_i s_i^2)$, where $\gamma_i = \frac{\sigma_v^2}{\sigma_v^2 + s_i^2}, i=1, \dots, m;$
- $[\beta | y, \theta, \sigma_v^2] \sim N_p \left(\left(\sum_{i=1}^m x_i x_i' \right)^{-1} \left(\sum_{i=1}^m x_i \theta_i \right), \sigma_v^2 \left(\sum_{i=1}^m x_i x_i' \right)^{-1} \right);$
- $[\sigma_v^2 | y, \theta, \beta] \sim \text{IG} \left(a_0 + \frac{m}{2}, b_0 + \frac{1}{2} \sum_{i=1}^m (\theta_i - x'_i \beta)^2 \right).$

Under Model 2, the HB estimator of θ_i and the corresponding posterior variance estimator are given by (4) and (5) respectively with $\sigma_i^{2(lg)}$ replaced by s_i^2 . Note that using s_i^2 instead of $\sigma_i^{2(lg)}$ may lead to severe underestimation of the posterior variance of θ_i for some areas with small sample sizes n_i . We will compare the HB estimators and evaluate the effects of using s_i^2 in Model 2 through data analysis in the following section.

3. Data Analysis

3.1 The Data Sets

We consider two interesting data sets in our analysis. The first data set is corn and soybean data with only 8 areas and small sample sizes in each area. The second data set is milk data with 43 areas and relatively large sample sizes in each area. We will compare the HB models and estimates based on these two data sets.

Corn and Soybean Data: The corn and soybean data comes from the U.S. Department of Agriculture and was first studied by Battese, Harter and Fuller (1988). The data contains reported crop hectares and LANDSAT satellite data for corn and soybeans in sample segments of 12 Iowa counties. The reported number of hectares for each crop comprise the direct survey estimates. Used as auxiliary data are the population means of number of pixels of a given

crop per segment. The sample sizes are small for these areas, ranging from 1–5. For our purposes only the counties with a sample size of 3 and greater are used (8 areas meet the criteria). Therefore of the included counties the sample sizes range from 3–5. The original data is unit level data. In order to have area level data the sample mean and the sample standard error are calculated for each county. The sample standard errors for the corn and soybean data are quite large in general (yielding some CVs in the 0.3–0.4 range and one CV of 0.532) but by chance there are also some small values in some instances (for corn data, Franklin has standard error 5.704 and CV 0.036). Because the sample sizes are so small, these sample standard errors cannot be trusted to approximate the true standard errors. Table 1 presents the modified area level data for corn and soybeans from the unit level data of Battese *et al.* (1988).

Table 1
Modified Crop Area Level Data, from
Battese, Harter and Fuller (1988)

County	n_i	Corn				Soybeans		
		y_i	SD	CV		y_i	SD	CV
Franklin	3	158.623	5.704	0.036		52.473	16.425	0.313
Pocahontas	3	102.523	43.406	0.423		118.697	50.290	0.424
Winnebago	3	112.773	30.547	0.271		88.573	10.453	0.118
Wright	3	144.297	53.999	0.374		97.800	52.034	0.532
Webster	4	117.595	21.298	0.181		112.980	23.531	0.208
Hancock	5	109.382	15.661	0.143		117.478	17.209	0.146
Kossuth	5	110.252	12.112	0.110		117.844	20.954	0.178
Hardin	5	120.054	36.807	0.307		101.834	26.790	0.263

Milk Data: The milk data, used in an article by Arora and Lahiri (1997), comes from the U.S. Bureau of Labor Statistics. The estimated values are the average expenditure on fresh milk for the year 1989. There is data for 43 areas with sample sizes ranging from 95 to 633. The CVs range from 0.074 to 0.341 over the 43 areas. A more detailed description of the data can be found in Arora and Lahiri (1997). For completeness, we give the data in Table 2. Following Arora and Lahiri (1997), we use $x'_i \beta = \beta_j$ if $i \in j^{\text{th}}$ major area, a collection of similar publication areas. Arora and Lahiri (1997) used eight major areas. Since this division of the eight major areas is not given in their paper, after noting trends in the data we used the Fay-Herriot model to test two new divisions of 6 and 4 major areas that combine similar survey estimates. These major areas produced large CV reduction in general. Where the 6 groups had yielded an average CV reduction of about 20% the 4 groups gave approximately an average 25% CV reduction over the direct estimates. Comparison of the point estimates and CVs have shown that the 4 major areas perform better than the 6 major areas. The 4 major areas are 1–7, 8–14, 15–25 and 26–43. In this paper, we will use these 4 groups as auxiliary variables for illustration purpose only.

Table 2
Milk Data, from Arora and Lahiri (1997)

Small Area	n_i	y_i	SD	CV
1	191	1.099	0.163	0.148
2	633	1.075	0.080	0.074
3	597	1.105	0.083	0.075
4	221	0.628	0.109	0.174
5	195	0.753	0.119	0.158
6	191	0.981	0.141	0.144
7	183	1.257	0.202	0.161
8	188	1.095	0.127	0.116
9	204	1.405	0.168	0.120
10	188	1.356	0.178	0.131
11	149	0.615	0.100	0.163
12	290	1.460	0.201	0.138
13	250	1.338	0.148	0.111
14	194	0.854	0.143	0.167
15	184	1.176	0.149	0.127
16	193	1.111	0.145	0.131
17	218	1.257	0.135	0.107
18	266	1.430	0.172	0.120
19	214	1.278	0.137	0.107
20	213	1.292	0.163	0.126
21	196	1.002	0.125	0.125
22	95	1.183	0.247	0.209
23	195	1.044	0.140	0.134
24	187	1.267	0.171	0.135
25	479	1.193	0.106	0.089
26	230	0.791	0.121	0.153
27	186	0.795	0.121	0.152
28	199	0.759	0.259	0.341
29	238	0.796	0.106	0.133
30	207	0.565	0.089	0.158
31	165	0.886	0.225	0.254
32	153	0.952	0.205	0.215
33	210	0.807	0.119	0.147
34	383	0.582	0.067	0.115
35	255	0.684	0.106	0.155
36	226	0.787	0.126	0.160
37	224	0.440	0.092	0.209
38	212	0.759	0.132	0.174
39	211	0.770	0.100	0.130
40	179	0.800	0.113	0.141
41	312	0.756	0.083	0.110
42	241	0.865	0.121	0.140
43	205	0.640	0.129	0.202

3.2 Analysis of Results

Corn and Soybean Data: First we consider the effect of our treatment of σ_i^2 using the HB approach. Table 3 presents the HB estimates $\hat{\theta}_i^{\text{HB}}$ and the associated standard errors (SDs) and CVs for the small area corn and soybean data sets. The SD is the square root of the posterior variance. Under Model 1 (σ_i^2 unknown), the SDs and CVs are consistently larger than the corresponding SDs and CVs under Model 2 ($\sigma_i^2 = s_i^2$ known). The increased SDs and CVs of Model 1 are expected since this model takes into account the added variability of estimating σ_i^2 . On average there is about 20% increase in SDs and CVs (this calculation excludes Franklin for corn data). The results support the fact that letting $\sigma_i^2 = s_i^2$, the known direct estimate of σ_i^2 , leads to underestimation of the SD and CV of $\hat{\theta}_i$.

Inspection of small areas Franklin and Webster for the corn data and county Winnebago for the soybean data establish in some cases where the sampling errors by chance are quite small this under estimation is severe.

Comparison of the HB estimates under Model 1 and Model 2 to the direct estimates can be made using the CVs in Table 1 and Table 3. Under Model 2 the HB estimates have smaller CVs than the direct estimates in 6 of the 8 counties for the corn data and similarly for the soybean data, 6 out of 8 counties. Of the remaining 2 counties for each crop, the CVs under Model 2 are the same as the direct survey CVs or only slightly larger. Estimators from Model 2 therefore seem to have gained efficiency compared to the direct survey estimators. Now examining the HB estimates under Model 1 and the direct survey estimates lead to mixed results for the corn and soybean data sets. Model 1 accounts for the added uncertainty of estimating the sampling variances and so in only 4 of the 8 counties the HB estimates show improvements in efficiency for the corn data. For the soybean data 5 out of 8 counties demonstrate the HB estimates as improvements on the direct survey CVs. For the remaining counties the direct estimates exhibit lower CVs and even substantially lower CVs in some cases. For the corn data, counties Franklin and Webster have CV increases with Model 1 of more than 0.09 and 0.12 respectively. As well for the soybean data, county Winnebago has a CV increase of almost 0.10 from the direct survey estimate, using Model 1. Areas where the direct estimates demonstrate smaller CVs compared to the HB estimates include a number of those areas where the CVs are by chance atypically small. So the increased model-based CVs may reflect more appropriate CVs for those areas. Of the 7 cases where the direct CVs are smaller compared to the HB CVs under Model 1, the 3 cases noted above have severe differences and the remaining 4 instances show only slight reduction in efficiency with use of Model 1. Since direct survey estimates quite often have unacceptably large CVs and yet still by chance may have CVs grossly and inexplicably small, HB estimation under Model 1 may be more reliable and reasonable by taking into consideration the uncertainty of estimating σ_i^2 .

Milk data: Table 4 contains the HB estimates for the milk data. As expected, over the 43 areas the treatment of σ_i^2 as known or unknown shows negligible differences in terms of point estimates, SDs and CVs due to the large sample sizes in the 43 areas. Therefore the substitution of $\sigma_i^2 = s_i^2$ in the model is reasonable when the area-specific sample sizes are large, as clearly shown in this example. Also the HB estimates give reduced SDs and CVs when compared to the direct survey estimates in Table 2. As would be expected, the HB estimation approach is thus an improvement on the direct survey estimates.

Table 3
Comparison of HB Estimates for Crop Data

County	σ_i^2 known ($\sigma_i^2 = s_i^2$)			σ_i^2 unknown		
	$\hat{\theta}_i^{HB}$	SD	CV	$\hat{\theta}_i^{HB}$	SD	CV
Corn						
Franklin	155.788	6.061	0.039	142.862	18.408	0.129
Pocahontas	100.813	28.297	0.281	91.560	32.420	0.356
Winnebago	115.337	28.406	0.246	113.130	35.207	0.311
Wright	131.630	28.345	0.215	123.547	30.764	0.250
Webster	109.030	20.634	0.189	97.856	29.834	0.307
Hancock	121.682	15.656	0.129	123.478	17.857	0.145
Kossuth	115.710	11.180	0.097	114.910	12.510	0.109
Hardin	135.626	23.228	0.171	135.178	23.804	0.176
Soybean						
Franklin	75.375	16.272	0.216	88.186	21.067	0.239
Pocahontas	116.943	27.031	0.231	109.052	30.098	0.276
Winnebago	87.525	10.304	0.118	88.053	18.854	0.214
Wright	104.184	23.671	0.227	105.825	24.497	0.232
Webster	115.510	20.789	0.180	109.455	25.801	0.236
Hancock	101.368	15.741	0.155	102.876	17.311	0.169
Kossuth	102.388	14.948	0.146	101.862	15.019	0.148
Hardin	87.455	17.774	0.203	93.397	20.251	0.217

Table 4
Comparison of HB Estimates for Milk Data

Small area	σ_i^2 known ($\sigma_i^2 = s_i^2$)			σ_i^2 unknown		
	$\hat{\theta}_i^{HB}$	SD	CV	$\hat{\theta}_i^{HB}$	SD	CV
1	1.020	0.113	0.111	1.021	0.111	0.109
2	1.045	0.072	0.069	1.045	0.071	0.068
3	1.065	0.073	0.069	1.065	0.074	0.069
4	0.767	0.095	0.124	0.770	0.096	0.125
5	0.849	0.096	0.113	0.852	0.096	0.113
6	0.975	0.103	0.106	0.975	0.102	0.105
7	1.058	0.125	0.118	1.055	0.125	0.118
8	1.097	0.099	0.090	1.096	0.099	0.090
9	1.219	0.121	0.099	1.215	0.121	0.100
10	1.192	0.122	0.102	1.190	0.122	0.102
11	0.793	0.094	0.119	0.799	0.097	0.122
12	1.213	0.131	0.108	1.209	0.130	0.107
13	1.206	0.112	0.093	1.203	0.112	0.093
14	0.984	0.107	0.109	0.987	0.107	0.109
15	1.187	0.105	0.088	1.187	0.104	0.087
16	1.156	0.104	0.090	1.156	0.102	0.089
17	1.225	0.101	0.083	1.225	0.100	0.081
18	1.284	0.115	0.089	1.281	0.113	0.088
19	1.234	0.101	0.082	1.235	0.100	0.081
20	1.233	0.110	0.089	1.233	0.110	0.089
21	1.092	0.097	0.089	1.095	0.098	0.089
22	1.192	0.128	0.107	1.193	0.127	0.106
23	1.122	0.103	0.092	1.125	0.103	0.091
24	1.221	0.113	0.092	1.220	0.111	0.091
25	1.193	0.086	0.072	1.193	0.086	0.072
26	0.761	0.091	0.120	0.762	0.091	0.120
27	0.763	0.092	0.120	0.762	0.091	0.119
28	0.734	0.125	0.170	0.732	0.123	0.169
29	0.768	0.085	0.110	0.767	0.085	0.110
30	0.615	0.076	0.124	0.618	0.076	0.123
31	0.769	0.122	0.158	0.767	0.120	0.156
32	0.795	0.119	0.150	0.792	0.118	0.148
33	0.771	0.091	0.118	0.770	0.090	0.117
34	0.612	0.060	0.099	0.613	0.062	0.100
35	0.701	0.085	0.121	0.701	0.084	0.120
36	0.757	0.094	0.123	0.759	0.093	0.123
37	0.534	0.080	0.150	0.538	0.081	0.151
38	0.744	0.096	0.129	0.743	0.095	0.128
39	0.754	0.082	0.108	0.753	0.082	0.108
40	0.768	0.088	0.115	0.768	0.088	0.115
41	0.747	0.071	0.095	0.747	0.070	0.094
42	0.801	0.093	0.116	0.800	0.092	0.116
43	0.682	0.094	0.139	0.682	0.094	0.138

3.3 Priors and Sensitivity Analysis

In Model 1, the sampling variances σ_i^2 are assumed to be independent with inverse gamma prior distribution $IG(a_i, b_i)$, and the model variance σ_v^2 also has inverse gamma prior distribution $IG(a_0, b_0)$, where a_i, b_i ($0 \leq i \leq m$) are chosen to be very small known constants to reflect vague knowledge on σ_i^2 and σ_v^2 . So we have used proper priors to avoid the problem of any improper posteriors. One may consider using flat priors for σ_i^2 and σ_v^2 , i.e., $\pi(\sigma_i^2) \propto 1$, and $\pi(\sigma_v^2) \propto 1$, similar to the flat prior on β . With the flat priors on σ_i^2 and σ_v^2 , the full conditional distributions for σ_i^2 and σ_v^2 are given as

$$[\sigma_i^2 | y, \theta, \beta, \sigma_v^2] \sim IG\left(\frac{d_i - 1}{2}, \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2}\right),$$

and

$$[\sigma_v^2 | y, \theta, \beta, \sigma_i^2] \sim IG\left(\frac{m - 2}{2}, \frac{1}{2} \sum_{i=1}^m (\theta_i - x'_i \beta)^2\right).$$

The implementation of the Gibbs sampler under the flat priors is also straightforward. However, the flat priors on σ_i^2 and σ_v^2 may lead to improper posteriors if the sample sizes and the number of small areas are small. In order to see the problem on σ_i^2 more clearly, we can study the Model 1 in two steps. First, we can obtain the posterior of σ_i^2 given its direct estimate s_i^2 as

$$\begin{aligned} \pi(\sigma_i^2 | s_i^2) &\propto f(s_i^2 | \sigma_i^2) \cdot \pi(\sigma_i^2) \\ &\propto (\sigma_i^2)^{-d_i/2} \cdot \exp\{-\sigma_i^{-2} d_i s_i^2 / 2\} \cdot \pi(\sigma_i^2). \end{aligned}$$

By assuming a flat prior $\pi(\sigma_i^2) \propto 1$, we can obtain

$$\pi(\sigma_i^2 | s_i^2) \sim IG\left(\frac{d_i}{2}, \frac{d_i s_i^2}{2}\right),$$

provided that $d_i > 2$, or $n_i > 3$. Then we can use this proper IG posterior $\pi(\sigma_i^2 | s_i^2)$ as an informative prior for σ_i^2 in the sampling model $y_i | \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2)$. This will ensure to have proper posterior inference. For the modified corn and soybean data, using flat priors on σ_i^2 will lead to improper posterior due to the small sample sizes ($n_i = 3$) for some areas. Thus, proper inverse gamma priors are used in the data analysis to ensure that all the posteriors are proper, as commonly used in the HB small area estimation in practice (e.g., Arora and Lahiri 1997; Datta, Lahiri, Maiti and Lu 1999; You and Rao 2000; Rao 2003). Hence we do not face the problem of some posteriors being improper, since correct HB inference should be based on proper posteriors. Under Model 2 with the sampling variance known as $\sigma_i^2 = s_i^2$, using a flat prior $\pi(\sigma_i^2) \propto 1$ for σ_v^2 , the posterior of σ_v^2 will be proper provided that

$m > p + 2$, where m is the number of small areas and p is the size of regression parameters β (Rao 2003, page 238). Since the number of small areas is usually relatively large, this condition is generally satisfied in practice.

For the sensitivity analysis of vague proper priors, we can test the sensitivity of the posterior estimates to the choice of prior parameters $a_i, b_i (0 \leq i \leq m)$. Under Model 1, we set $a_i = b_i$ at four different values, i.e., 0.0001, 0.001, 0.01 and 0.1. Table 5 presents the estimated posterior means for the corn and soybean data, and Table 6 presents the corresponding CVs.

Table 5
Comparison of Posterior Mean Estimates for Crop Data

County	IG (a_i, b_i), $a_i = b_i$			
	0.0001	0.001	0.01	0.1
Corn				
Franklin	142.862	142.593	143.155	144.311
Pocahontas	91.560	91.912	91.422	91.974
Winnebago	113.130	113.068	121.578	114.430
Wright	123.547	124.170	125.103	125.351
Webster	97.856	98.231	99.132	98.511
Hancock	123.478	123.858	124.395	124.138
Kossuth	114.910	115.281	115.316	115.528
Hardin	135.178	134.157	135.223	136.001
Soybean				
Franklin	88.186	89.368	89.145	89.513
Pocahontas	109.052	109.571	107.745	108.176
Winnebago	88.053	87.478	86.267	87.302
Wright	105.825	106.712	105.142	104.676
Webster	109.455	108.392	109.835	110.252
Hancock	102.876	103.413	102.240	101.808
Kossuth	101.862	101.159	101.379	100.808
Hardin	93.397	94.713	93.576	94.767

Table 6
Comparison of Posterior CVs for Crop Data

County	IG (a_i, b_i), $a_i = b_i$			
	0.0001	0.001	0.01	0.1
Corn				
Franklin	0.129	0.124	0.128	0.125
Pocahontas	0.356	0.351	0.347	0.341
Winnebago	0.311	0.314	0.321	0.324
Wright	0.250	0.246	0.235	0.236
Webster	0.307	0.292	0.285	0.280
Hancock	0.145	0.148	0.148	0.142
Kossuth	0.109	0.110	0.107	0.104
Hardin	0.176	0.173	0.178	0.168
Soybean				
Franklin	0.239	0.233	0.231	0.227
Pocahontas	0.276	0.281	0.271	0.296
Winnebago	0.214	0.193	0.196	0.198
Wright	0.232	0.223	0.231	0.226
Webster	0.236	0.231	0.237	0.228
Hancock	0.169	0.165	0.168	0.161
Kossuth	0.148	0.145	0.142	0.135
Hardin	0.217	0.215	0.213	0.213

It is clear from Table 5 and Table 6 that the posterior estimates and the corresponding CVs are about the same and stable, which indicates that the HB estimates are not

sensitive to the choice of vague proper priors. For the milk data, the HB estimates are very stable to these proper vague priors (results are not provided here). Since the milk data has large sample sizes, flat priors on variance components can also be used to analyze the milk data under Model 1. We thus obtained the HB estimates based on the flat priors and compared them with the HB estimates based on the vague IG priors. These HB estimates are almost identical and stable with relative difference ranging from 0.07% to 2.23%, an average value of 0.69% over 43 areas, which indicates that the posterior estimates of small area means based on Model 1 are very stable and not sensitive to the choice of flat priors or vague IG priors, provided that the sample sizes and number of small areas are relatively large.

4. Conclusion and Future Work

In this paper we have studied the well-known Fay-Herriot model with the situations where σ_i^2 , the sampling error variances, are assumed unknown and where they are estimated by unbiased estimators s_i^2 , using the HB approach. The full HB approach with the Gibbs sampling method automatically takes into account the extra uncertainty associated with the estimation of σ_i^2 . We applied the HB approach in two survey data analysis. Our results have shown that the proposed HB approach under Model 1 works quite well no matter the area-specific sample sizes are small or large. For future work, the proposed HB modeling approach can be extended to the general area level models studied by You and Rao (2002). Application of the new HB modeling approach includes the census undercoverage estimation as in You, Rao and Dick (2004). Under Model 1, the HB estimators of the sampling variances σ_i^2 can be obtained. These HB estimators of σ_i^2 can be used as alternative smoothed estimators for σ_i^2 in the sampling models. Application and evaluation of the HB estimators of the sampling variances include the census undercoverage estimation and the Canadian Labour Force Survey (LFS) unemployment rate estimation (You, Rao and Gambino 2003). We also plan to compare the HB approach with the EBLUP approach as studied by Rivest and Vandal (2002) and Wang and Fuller (2003).

Acknowledgements

The authors would like to thank two referees, an Associate Editor, the Deputy Editor and the Editor, Dr. M.P. Singh, for their constructive comments and suggestions. The authors also would like to thank J.N.K. Rao of Carleton University for his useful suggestion and Jack Gambino and Eric Rancourt of Statistics Canada

for their comments on the early version of the paper. This work was supported by Statistics Canada Methodology Branch Research Block Fund.

References

- Arora, V., and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.
- Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999) Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.
- Fay, R.E., and Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.
- Gelfand, A.E., and Smith, A.F.M. (1990). Sample-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 972-985.
- Gelfand, A.E., and Smith, A.F.M. (1991). Gibbs sampling for marginal posterior expectations. *Communications In Statistics – Theory and Methods*, 20, 1747-1766.
- Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Rivest, L.P., and Vandal, N. (2002). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, July 10-13, 2002, Ottawa, Canada.
- Wang, J., and Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.
- You, Y., and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, 26, 173-181.
- You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 1, 3-15.
- You, Y., Rao, J.N.K. and Dick, P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.
- You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 25-32.

A Cost-Effective Strategy for Provincial Unemployment Estimation: A Small Area Approach

Ali-Reza Khoshgooyanfar and Mohammad Taheri Monazzah¹

Abstract

This paper primarily aims at proposing a cost-effective strategy to estimate the intercensal unemployment rate at the provincial level in Iran. Taking advantage of the small area estimation (SAE) methods, this strategy is based on a single sampling at the national level. Three methods of synthetic, composite, and empirical Bayes estimators are used to find the indirect estimates of interest for the year 1996. Findings not only confirm the adequacy of the suggested strategy, but they also indicate that the composite and empirical Bayes estimators perform well and similarly.

Key Words: Composite estimator; Design-based estimator; Empirical Bayes estimator; Indirect estimator; Non-sampling error; Synthetic estimator; Post-strata.

1. Introduction

Each year, sample surveys are conducted in Iran to obtain statistical information required for decision and policy making. However, these surveys cannot fulfill all statistical requirements because of two factors. The first one is related to the governmental and non-governmental sectors' demand for comprehensive statistical information not only at national and regional but also at small area levels. Further, they need the information at shorter periods of time per year, say monthly or quarterly. The second factor is that the main source of statistical data in Iran is surveys, and there are financial limitations for conducting surveys several times per year at small area levels. These two factors challenge statistical agencies to find efficient strategies to balance both cost and statistical information quality. The work presented here is an endeavor to overcome this challenge by using small area estimation (SAE) methods.

The purpose of SAE methods is to provide acceptable estimates for some subpopulations in a sample design planned for the "whole" population regardless of the subpopulations. For example, a sample design is planned for estimating population parameters for the "country" and after data collection the parameters are estimated by the national sample data. If simultaneously "provincial estimates" of the parameters are needed, it is not possible to conduct separate provincial sample surveys. The provinces are unplanned subpopulations in the sense that the available sample design has been planned just for estimating the parameters for the country without considering the provincial level. In the nationwide sample, few or no sample units may be available for some provinces. Hence, acceptable estimates for such provinces (subpopulations) cannot be produced.

Before the availability of SAE methods, such subpopulation estimates were obtained by direct design-based estimation. If there were data from a given subpopulation in the nationwide sample, an estimate would be directly calculated according to the nationwide sample design by using "the available data". The direct estimate may differ substantially from the actual subpopulation parameter due to large sampling errors owing to small sample size.

Statisticians and demographers have developed ways of estimating for such subpopulations. Indirect estimators have been suggested and applications have been increasing over the last twenty years. However, the SAE methods are still an active topic of study. See Purcell and Kish (1979, 1980), Ghosh and Rao (1994), Schaible (1995), Marker (1999), Pfeffermann (2002) and especially Rao (2003a) for problem definition and a review of the SAE methods.

For a number of years, the Statistical Center of Iran (SCI) implemented annually a national one-stage cluster sample in order to estimate the intercensal unemployment rate at the country level. For sixteen years, separate one-stage cluster samples for all provinces have been conducted to estimate provincial unemployment rates. A weighted combination of provincial estimates then yields the unemployment rate for the total country. The increasing need for estimation of the unemployment rate at a provincial level on a monthly, or at least seasonal basis, and the lack of administrative records in Iran at both small and national levels persuaded SCI to try the SAE methods as the core of a revised strategy to meet the provincial need.

The revised strategy consists of designing a sample survey only at the national level and producing the provincial estimates by SAE methods. A province in the strategy is a small area. This strategy demands a smaller sample size than that for aggregating provincial samples. If the revised

1. Ali-Reza Khoshgooyanfar, The Center for Research, Studies and Program Assessments of IRIB. E-mail: khosh_ar@yahoo.com; Mohammad Taheri Monazzah, The Central Bank of Iran. E-mail: Taheri53@yahoo.com.

strategy proves practicable, time and cost of the data collection can be reduced, and produce provincial estimates on a monthly basis. The smaller sample is easier to control in the field, and estimates are less affected by nonsampling errors.

This paper is intended to answer the following questions:

1. Can a nationwide sample substitute for separate provincial samples for making estimates of the provincial unemployment rates?
2. From the three SAE methods – synthetic, composite and empirical Bayes estimators – which one produces the best estimates?

To answer empirically these two questions, estimates were produced for the year 1996 when the actual values of the provincial unemployment rates are available from the 1996 Census. Consequently, the actual bias of each provincial estimate can be computed.

The process includes the following three stages. First, a sample of size 13,000 from the whole country is selected (the 1996 Census data file). The sample size is determined at the national level, and is allocated to all provinces proportionally to population. The allocation provides sample from each province enabling direct estimates of the unemployment rate for each province. Direct estimates are not necessarily acceptable for all provinces because of the large sampling errors due to small sample sizes in some provinces. Second, applying three SAE methods, indirect estimates are produced for each province. Third, the indirect estimates are evaluated by comparing them with corresponding actual values, computing MSEs, mean of absolute errors (MAE), and mean of errors (ME).

In addition to this introduction, the paper takes in three more sections. Section 2 offers a short review of the three estimators used in this paper, including the estimation methods, their corresponding MSEs, and properties of the estimators. The estimates and corresponding computational aspects are presented in section 3, where performances of the estimators are tentatively appraised. Section 4 is devoted to final remarks and recommendations about the estimators and the merit of the SAE strategy.

2. A Glance Over the Estimators

Indirect estimators used in the study are introduced briefly. However, an excellent discussion of the SAE methodology is in Rao (2003a). First, the synthetic estimator is considered, and then the composite estimator. The empirical Bayes (EB) estimator as a model-based estimator is also considered.

2.1 Synthetic Estimator

There is a family of small area estimators characterized as synthetic, see Rao (2003a, chapter 4). The traditional and simplest is discussed here. For this estimator,

1. The country is partitioned into six post-strata on the basis of six age groups, see Table (1).
2. Next, the number of unemployed persons is estimated in each province, providing the numerator in expression (1).
3. The synthetic estimate of the i^{th} province is obtained by dividing the estimated number of unemployed persons in province i by its Economically Active Population (EAP), namely

$$\hat{P}_i^S = \left(\sum_{j=1}^6 N_{ij} \hat{P}_j \right) / N_i \quad (1)$$

where \hat{P}_j is a direct design-based estimate of the unemployment rate in post-stratum j , N_i is the EAP in province i , and N_{ij} is the EAP in the intersection of province i and post-stratum j , cell (i, j) . The synthetic estimate of the i^{th} province is according to the official definition of the unemployment rate in Iran.

The synthetic estimate shares all national sample data by using national direct estimates of the unemployment rate from the post-strata. It uses the six estimated “post-strata” unemployment rates computed over all provinces rather than specific estimates of the six “cells”. This process thus **borrows strength** because each province contributes to the national sample by pooling provincial sample units to overcome small sample sizes in each province.

This estimator has three limitations:

1. The smaller the inter-post-stratum variation is, the better synthetic estimator performs. It means that all provinces should have a rather equal unemployment rate in each age group. Using the national post-strata direct estimates equally for all provinces is allowable only under this assumption. If the homogeneity assumption is not satisfied, the synthetic estimator cannot reflect specific small area variation, and the estimates could be severely biased.
2. If there are several variables that are important in post-stratification, the synthetic estimator cannot often use all of them because post-strata (after cross-classification of the several variables) have sample sizes that are too small and yield unacceptable direct estimates of the post-strata. Generally speaking, many post-strata give rise to poor direct estimates for some of the post-strata. This can create serious problems for synthetic estimation when a poor direct estimate receives a large EAP for a cell.

3. Quality of the EAPs can affect the synthetic estimates. Owing to lack of timely data sources such as administrative records, out of date EAPs from the 1986 Census data are used here in order to produce the synthetic estimates for the year 1996.

2.2 Composite Estimator

The composite estimator of the i^{th} province combines the synthetic and direct estimators of that province, namely

$$\hat{P}_i^C = W_i \hat{P}_i^D + (1 - W_i) \hat{P}_i^S \quad (2)$$

where \hat{P}_i^D is the direct design-based estimator for the i^{th} province, and $0 \leq W_i \leq 1$. Expression (2) improves upon (1) by exploiting both estimators. That is, provincial differences may take into account in the composite estimator via the provincial unbiased direct estimates and instability of the direct estimator may be reduced via the synthetic estimator.

The weight W_i can be specified so as to minimize mean square error of \hat{P}_i^C , $\text{MSE}(\hat{P}_i^C)$. Assuming $\text{Cov}(\hat{P}_i^D, \hat{P}_i^S) \cong 0$, the weight is simplified as

$$W_i^{\text{opt}} = \frac{1}{(V(\hat{P}_i^D)/\text{MSE}(\hat{P}_i^S)) + 1} \quad (3)$$

where $V(\hat{P}_i^D)$ and $\text{MSE}(\hat{P}_i^S)$ are the variance of \hat{P}_i^D and the mean square error of \hat{P}_i^S , respectively. In expression (3), the weights of the direct and synthetic estimators in (2) are proportional to the MSEs of the two estimators. See Schaible (1978) and Rao (2003a, page 58) for properties of the estimator and weight.

In practice, we should estimate $\text{MSE}(\hat{P}_i^S)$ and $V(\hat{P}_i^D)$ to generate an estimate of the weight (3). If there are some sample data from the i^{th} province, according to the sample design, an unbiased design-based estimator of $V(\hat{P}_i^D)$ can be computed by using only the sample data. Therefore, only an estimator for $\text{MSE}(\hat{P}_i^S)$ is required. Under the assumption that $\text{Cov}(\hat{P}_i^D, \hat{P}_i^S) \cong 0$, Ghosh and Rao (1994) proposed the unbiased estimator

$$\hat{\text{MSE}}(\hat{P}_i^S) = (\hat{P}_i^S - \hat{P}_i^D)^2 - \hat{V}(\hat{P}_i^D). \quad (4)$$

Under the same assumption, one can easily show that

$$\text{MSE}(\hat{P}_i^C) = W_i^2 V(\hat{P}_i^D) + (1 - W_i)^2 \text{MSE}(\hat{P}_i^S). \quad (5)$$

The estimator (4) may result in negative estimates for some provinces, and the weight in expression (3) is no longer computable. In this case, instead of (3) and (4), we have used respectively the combined weight in (6) and $\hat{\text{AMSE}} = (1/I') \sum_{i=1}^{I'} \hat{\text{MSE}}(\hat{P}_i^S)$ where I' is the number of small areas having positive estimated MSE (see Gonzalez and Waksberg (1973) for more details):

$$W^C = \frac{1}{\left(\sum_i \hat{V}(\hat{P}_i^D) / \sum_i \hat{\text{MSE}}(\hat{P}_i^S) \right) + 1}. \quad (6)$$

In addition to expressions (3) and (6), Copas (1972), Ghosh and Rao (1994), and Thompson and Holmoy (1998) suggest alternative weights.

2.3 Empirical Bayes (EB) Estimator

Model based SAE methods have received more attention than the synthetic and composite estimators. Marker (1999) regarded the SAE methods as having a common element expressed through regression models. The EB method is of the regression type. Consider the following mixed model (see Rao (2003a, page 76)):

$$g = X\beta + \mathbf{v} + \varepsilon \quad (7)$$

where

$$\mathbf{g}' = (Ln \frac{\hat{P}_1^D}{1 - \hat{P}_1^D}, \dots, Ln \frac{\hat{P}_I^D}{1 - \hat{P}_I^D}),$$

X is an $I \times k$ design matrix of supplementary variables, β is a $k \times 1$ vector of unknown parameters, and \mathbf{v} and ε are $I \times 1$ random vectors (I is the number of provinces). Assume that:

1. \mathbf{v} and ε are independent.
2. $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \text{Diag}(d_1^2, \dots, d_I^2)$.
3. $\mathbf{v} \sim N(0, \Sigma)$ where $\Sigma = \text{Diag}(t^2, \dots, t^2)$.

Ghosh and Meeden (1997) show that the EB estimate of the i^{th} element of \mathbf{g} is:

$$\hat{g}_i^{\text{EB}} = \hat{W}_i \mathbf{x}_i' \hat{\beta} + (1 - \hat{W}_i) g_i \quad (8)$$

where \mathbf{x}_i' and g_i are the i^{th} row and the i^{th} component of X and \mathbf{g} respectively, and \hat{W}_i is an estimate of

$$W_i = \frac{d_i^2}{d_i^2 + t^2}. \quad (9)$$

Consequently, the EB estimate of the i^{th} rate is:

$$\hat{P}_i^{\text{EB}} = \frac{\exp(\hat{W}_i \mathbf{x}_i' \hat{\beta} + (1 - \hat{W}_i) g_i)}{1 + \exp(\hat{W}_i \mathbf{x}_i' \hat{\beta} + (1 - \hat{W}_i) g_i)}. \quad (10)$$

It is obvious that (10) needs two estimates for β and the weight in (9). On the other hand, the weight in (9) relies on the estimates of t^2 and d_i^2 . By applying the delta method, $(g_i')^2 \hat{V}(\hat{P}_i^D)$ generates an estimate of d_i^2 where g_i' through the first derivative of $g_i = Ln(\hat{P}_i^D / (1 - \hat{P}_i^D))$. Based on Chand and Alexander (1995), estimates of β and t^2 are found by simultaneously solving

$$\begin{cases} t^2 = (g - X\beta)'V^{-1}(g - X\beta)/(I - k) \\ \beta = (X'V^{-1}X)^{-1}X'V^{-1}g \end{cases} \quad (11)$$

where $V = \text{Diag}(d_1^2 + t^2, \dots, d_I^2 + t^2)$. Note that the equations in (11) are solved by numerical iteration with an initial value for t^2 .

The EB and composite estimators have similarities although they arise from different approaches. Both estimators have two components; a direct component (\hat{P}_i^D in (2) and g_i in (8)) computed from the provincial sample data, and an indirect component (\hat{P}_i^S in (2) and $x_i'\hat{\beta}$ in (8)) constructed from the national sample data and supplementary information. Both estimators (2) and (8) give more weight to the indirect component when it is reliable. Otherwise the direct component receives more weight. Additional details are given in Cressie (1989), Ghosh *et al.* (1998) and Rao (2003 a, b).

3. Estimation for Iran

Estimates were produced for the year 1996 because the 1996 actual unemployment rate of each province is known from the 1996 Census. As a result, the actual bias of each estimate can be computed.

In 1996 the country consisted of 26 provinces. However, 21 provinces are studied here because supplementary information from the 1986 Census was available for 21 geographically unchanged provinces between the years 1986 and 1996. To make the three indirect estimates, at the national level, a sample was planned and its sample size was determined for estimating the unemployment rate of the country as a whole. Each province is a small area. The national sample was allocated among the 26 provinces proportional to population in order to have sample data from each province (a top-down approach). This enabled direct design-based estimates for each province and its corresponding variance required for both the EB and composite estimators. The sample design is able to produce good estimates for the country and for some provinces.

3.1 Computational Aspects

To construct synthetic estimates, six age groups formed the post-strata. The estimated unemployment rate of each group based on the national sample and its corresponding actual value based on the 1996 Census are presented in Table (1), which also contains absolute errors of the estimates.

The estimates for the first two groups have very large error. Therefore, if a province in expression (1) gives large EAPs to these age groups, its synthetic estimate may not

perform well. The 1986 Census data were used in computing the EAPs for all provinces and cells (N_i and N_{ij} in expression (1)) because, in the absence of administrative records, the nearest census to the year 1996 is the main source of data at any level.

Table 1
Post-strata Characteristics

Age Group	Estimated Rate (\hat{P}_i)	Actual Value	Absolute Error
10–15	0.3240	0.2826	0.0414
16–20	0.2402	0.2629	0.0227
21–25	0.1868	0.1856	0.0012
26–30	0.0811	0.0802	0.0009
31–50	0.0363	0.0366	0.0003
More than 50	0.0653	0.0648	0.0005

To construct the composite estimates, provinces were divided into two groups. The first consists of 14 provinces that used the weight in expression (3), and the second of seven provinces used the common weight $W^C = 0.873184$ based on (6). Because the estimator in expression (4) produces negative estimates for $\text{MSE}(\hat{P}_i^S)$ for these seven provinces, the AMSE was used.

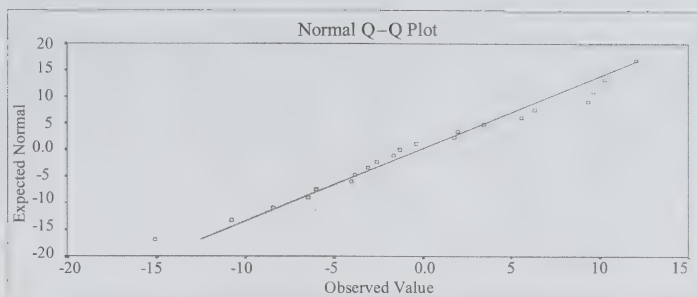
To construct the EB estimates, d_i^2 was estimated by using the delta method and then t^2 was estimated following Prasad and Rao (1990) by using a SAS/IML program (the program is available from the authors). An initial estimate for t^2 is required in this program, and was calculated by the moment estimation method as $t^2 = 0.3117194$. To solve the equations in (11), the following 21×2 design matrix was used, whose first and second columns are 1s and EAPs, respectively:

$$X = \begin{bmatrix} 1 & 133,449 \\ 1 & 141,124 \\ 1 & 883,653 \\ 1 & 795,714 \\ \vdots & \vdots \\ 1 & 522,976 \\ 1 & 162,892 \end{bmatrix}$$

The estimated t^2 and $\hat{\beta}$ are

$$\hat{t}^2 = 0.5596389, \hat{\beta} = \begin{pmatrix} -2.066874 \\ -1.273 \times 10^{-7} \end{pmatrix}$$

To test normality, a normal Q–Q plot and a Shapiro-Wilk's test for standardized residuals of the fitted model were examined. The points in the Q–Q plot are close to a straight line, and the test did not reject the null hypothesis of normality (p -value = 0.851).



3.2 Results

The results are organized into four parts. First, bias in the forms of error and absolute error is examined using two criteria, ME and MAE. Second, MSEs are compared among methods. Third, efficiencies of the indirect estimators relative to the direct estimator are evaluated. Finally, the weights of the direct components in expressions (2) and (8) are analyzed. All the results are depicted in appropriate figures, however, details can be found in Table (2).

Suppose S_a is the allocated sample size from the national sample to a given province and S_r the required sample size if the sample size is separately determined for the province. In other words, if there is a sample of size S_r from the province, an acceptable direct estimate can be then computed for the province. Therefore, $(S_a/S_r) \times 100$ measures how much the available sample size (S_a) is adequate for a given province. This measure is used on horizontal axes of all plots as a basis for comparison sample size effects.

The synthetic estimator has the highest MAE, which was even larger than that of the direct estimator (see Figure 1). Conversely, MAEs of the composite and EB estimators are the lowest, and very similar to one another. Based on ME, there is a slight overestimation of the actual value for all estimators. The direct estimator has the lowest ME because it is unbiased. MEs of the composite and EB estimators are close and the synthetic estimator has the highest ME.

For the direct, composite, and EB estimators, all provinces with $S_a/S_r \geq 10\%$ have absolute errors less than 0.02. The highest absolute errors belong to Ilam and Kohgiluyeh & Boyerahmad which have the smallest populations and very small S_a/S_r . Plots of these three estimators have relatively similar patterns. The story is different for the synthetic estimator because the "national" sample data are only used in making synthetic estimates through the post-strata direct estimates and then the "national" sample size (not S_a/S_r) affects the synthetic estimate of a province through the cell EAPs. In other words, if a post-stratum does not have "enough" national

sample data to yield acceptable direct estimates, and a province gives large EAP to the post-stratum direct estimate, the province has a poor synthetic estimate. This is the case for Sistan & Baluchestan, Bushehr, Tehran and Lorestan because of poor direct estimates for the first two post-strata (the age groups of 10–15 and 16–20) and large young populations of these provinces.

The lowest MSE always belongs to the composite or EB estimator (see Figure 2). However, MSE of the composite estimator is often lower than that of the EB estimator. The MSE of the synthetic estimator is always higher than those of the other estimators, even the direct estimator.

As the S_a/S_r increases, the MSE decreases for the direct, composite and EB estimators (see the descending trend in Figure 2). This effect is very drastic for Tehran ($S_a/S_r = 36\%$). Again, there are two exceptions for the three estimators, Ilam and Kohgiluyeh & Boyerahmad, both having the smallest populations and very small S_a/S_r . The pattern of Figure 2 for the synthetic estimator may be misleading because seven provinces used AMSE. However, the four previous provinces (Sistan & Baluchestan, Bushehr, Tehran and Lorestan) also do not conform to the pattern. As a general rule for an estimator, the greater the dependency on the provincial direct estimates the stronger the relationship between the MSE and the ratio S_a/S_r .

The relative efficiencies (RE) of the three indirect estimators compared to the direct estimator for all provinces are often smaller than or equal to one for the composite and EB estimators and greater than one for the synthetic estimator. Some composite estimates have good REs: Semnan (0.34), West Azarbayegan (0.46), Khorasan (0.70), Kermanshah (0.75) and Hamadan (0.77). Means of REs ($\overline{RE}^S = 13.6$, $\overline{RE}^C = 0.8595$ and $\overline{RE}^{EB} = 0.9951$) indicate that the composite estimator is the most efficient estimator among the three indirect estimators. Further, in Figure 3 as S_a/S_r increases RE^{EB} approaches one. Figure 3 as well as Figure 2 may be misleading for the synthetic estimator.

Table 2
Provincial and Estimator Characteristics

Province	EAP	S _a	S _r	S _a /S _r	RE ^C	RE ^{EB}	RE ^S	AE ^C	AE ^{EB}	AE ^S	AE ^D	MSE ^C	MSE ^{EB}	MSE ^S	MSE ^D
Bushehr	133,449	146	4,550	3.2%	0.96	1.17	25.57	0.03300	0.01687	0.06501	0.02644	0.0003030	0.000368	0.0080483	0.0003148
Chaharmahal & Bakhtiari*	141,124	203	4,063	5%	0.87	0.95	6.52	0.02136	0.02135	0.03644	0.02031	0.0003813	0.000417	0.0028670	0.0004397
Esfahan	883,653	1032	5,850	17.6%	0.90	1.00	9.56	0.01268	0.01421	0.00990	0.01504	0.0000533	0.000059	0.0005631	0.0000589
Fars	795,714	925	6,175	15%	0.91	0.99	9.69	0.00610	0.00886	0.02235	0.00904	0.0000836	0.000091	0.0008931	0.0000922
Gilan*	734,196	683	5,364	12.7%	1.04	0.97	17.25	0.00484	0.00460	0.01107	0.00393	0.0001728	0.000162	0.0028670	0.0001662
Hamadan	387,517	439	4,550	9.6%	0.77	1.00	3.36	0.01294	0.01701	0.00675	0.01880	0.0001155	0.000150	0.0005030	0.0001498
Hormozgan*	168,268	198	4,063	4.9%	0.84	0.93	5.12	0.01984	0.01734	0.02821	0.01862	0.0004731	0.000519	0.0028670	0.0005600
Ilam	84,210	111	4,063	2.7%	0.83	0.87	4.94	0.04901	0.05201	0.03395	0.06579	0.0013919	0.001450	0.0082747	0.0016734
Kerman*	312,768	450	5,200	8.7%	0.96	0.97	12.00	0.03615	0.03672	0.02864	0.03724	0.0002283	0.000231	0.0028670	0.0002389
Kermanshah	357,096	436	3,575	12.2%	0.75	0.96	3.07	0.00265	0.00928	0.02641	0.01210	0.0002747	0.000349	0.0011190	0.0003640
Khorasan	1,410,863	1,587	8,125	19.5%	0.70	0.99	2.36	0.00515	0.00193	0.01353	0.00160	0.0000298	0.000042	0.0000999	0.0000424
Khuzestan*	609,044	786	4,225	18.6%	1.03	0.97	16.83	0.01034	0.01247	0.00308	0.01140	0.0001760	0.000166	0.0028670	0.0001704
Kohgiluyeh & Boyerahmad	90,655	105	3,575	2.9%	0.83	0.86	4.83	0.05486	0.05932	0.02630	0.07165	0.0013629	0.001408	0.0079493	0.0016449
Kordestan*	276,575	341	5,200	6.6%	0.91	0.95	9.22	0.03105	0.02814	0.03641	0.03027	0.0002833	0.000297	0.0028670	0.0003111
Lorestan	310,918	341	3,575	9.5%	0.86	0.95	6.22	0.00943	0.01383	0.04101	0.01754	0.0004090	0.000451	0.0029534	0.0004747
Mazandaran*	917,259	1,043	6,013	17.3%	1.30	0.98	33.57	0.00199	0.00188	0.00310	0.00183	0.0001112	0.000084	0.0028670	0.0000854
Semnan	110,166	121	4,713	2.6%	0.34	1.08	0.51	0.02776	0.01929	0.03661	0.01042	0.0001534	0.000491	0.0002317	0.0004542
Sistan & Baluchestan	272,752	318	4,875	6.5%	0.96	0.97	26.53	0.00431	0.00228	0.08606	0.00123	0.0002519	0.000254	0.0069347	0.0002614
Tehran	2,343,290	2,913	8,125	35.9%	0.99	1.00	83.08	0.00605	0.00573	0.04767	0.00555	0.0000209	0.000021	0.0017530	0.0000211
West Azarbajejan	522,976	654	6,500	10.1%	0.46	0.98	0.85	0.00505	0.01247	0.00182	0.01309	0.0000552	0.000118	0.0001024	0.0001199
Yazd	162,892	207	5,038	4.1%	0.82	1.36	4.52	0.01414	0.00968	0.01008	0.01950	0.0001299	0.000215	0.0007164	0.0001586

*Denote provinces for which expression (3) produces negative estimates for MSEs.
EAP: Economically Active Population
S_a: Allocated Sample Size
S_r: Required Sample Size
RE: Relative Efficiency
AE: Absolute Error
MSE: Mean Squared Error (the lowest MSE is bold for each province)
C, EB, S and D stand for Composite, Empirical Bayes, Synthetic and Direct estimators, respectively.

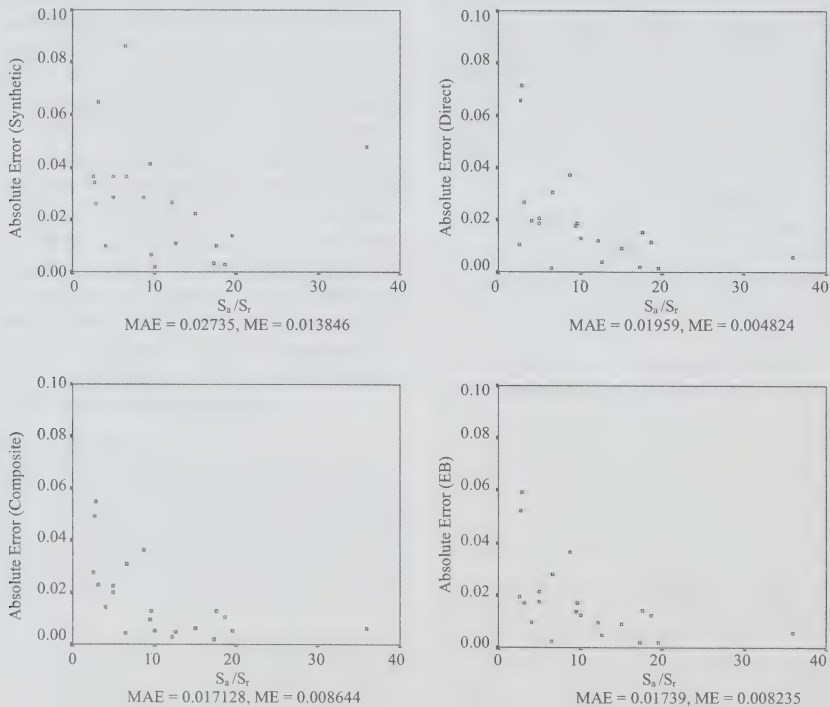


Figure 1. Absolute errors of the estimates against S_a/S_r .

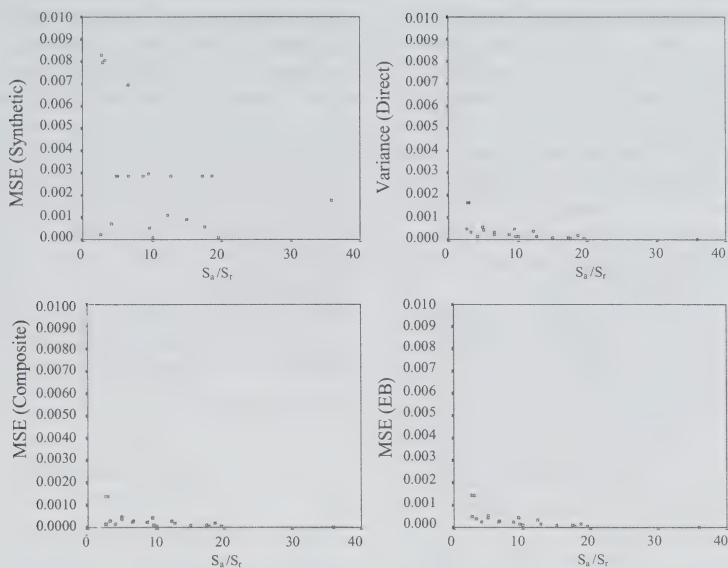


Figure 2. MSEs of the estimates against S_d/S_r .

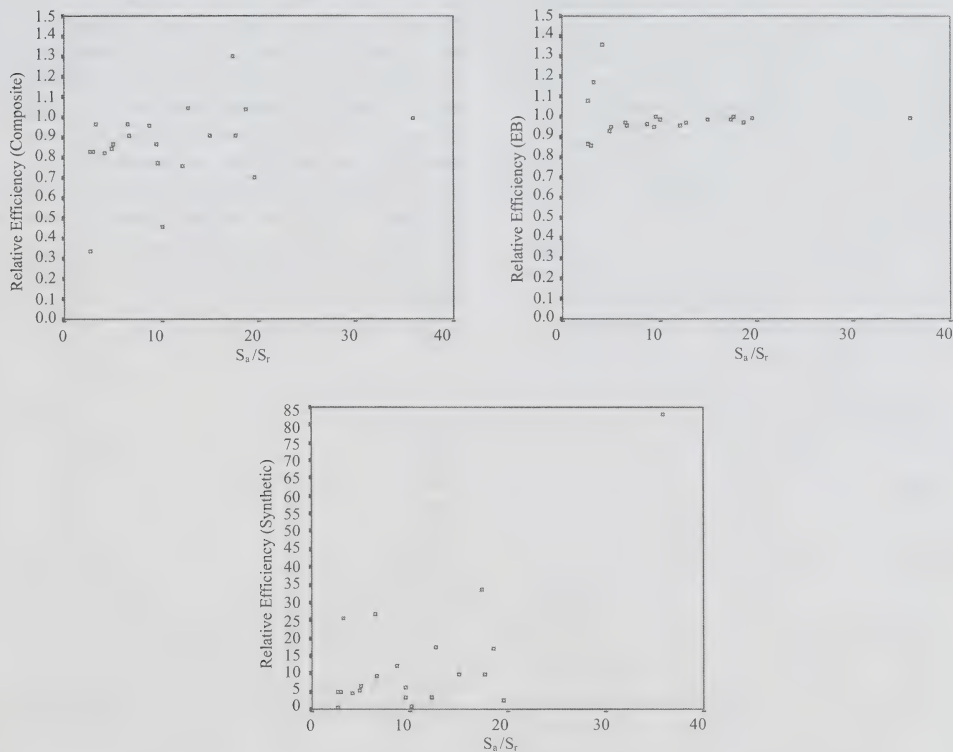


Figure 3. Relative efficiency (Estimated MSE of indirect estimator/Estimated variance of direct estimator) against S_d/S_r (the plot of synthetic estimator has a different scale on the vertical axis for legibility).

The direct component of the estimator in expression (8), g_1 , always receives more weight than the indirect component. This is the case for the composite estimator, except for two provinces of Semnan and West Azarbayejan. For the composite estimator, Rao (2003a, page 58) states that “the optimal weight W_i^{opt} will be close to zero or one when one of the component estimators has a much larger MSE than the other, that is, when $f_i = \text{MSE}(\hat{P}_i^C) / \text{MSE}(\hat{P}_i^S)$ is either large or small. In this case, the estimator with larger MSE adds little information and therefore it is better to use the component with smaller MSE.” This comment is clearly illustrated for Bushehr ($W = 0.962355$, $\text{RE}^S = 25.27$), Sistan & Baluchestan ($W = 0.963670$, $\text{RE}^S = 26.53$) and Tehran ($W = 0.988083$, $\text{RE}^S = 83.08$), because the direct estimates of these provinces have smaller MSEs than the synthetic estimates. Figure 4 clearly shows an ascendant relationship between the weight and S_a/S_r for the EB estimator. For the composite estimator, the lowest and highest weights pertain to the provinces with the lowest S_a/S_r and the highest S_a/S_r , respectively.

In general, the synthetic estimator performs poorly based on the MAE, ME, MSE and RE criteria, even though the synthetic estimates of some provinces are individually closer to actual values than other estimates. However, the synthetic estimates have been computed under the most disadvantageous conditions. The EAPs applied to construct the synthetic estimates are based on the 1986 Census (ten years before the year when the estimates were produced). In addition, the direct estimates of the first two post-strata are quite different from the other post-strata, causing poor synthetic estimates.

To address the first problem, administrative records should be developed; for the second, post-strata estimation should be handled in the sample design in advance. If not only post-strata estimation but also provincial classifications

in planning the sample design are considered in advance, good direct estimates for the post-strata can be expected. Consequently, good synthetic estimates for the provinces can be expected. The provincial classifications can increase homogeneity by putting similar provinces in classes together and using only sample data of the classified provinces to make the direct post-strata estimates to construct the synthetic estimates of those provinces.

The composite and EB estimators usually perform well when S_a/S_r is 10% or larger for a province because the direct components of the estimators (2) and (8) are relatively stable and receive a larger weight, especially for the EB estimator. Tehran, Khorasan, Khuzestan and Esfahan are of this type, while Bushehr, Ilam, Kohgiluyeh & Boyerahmad and Semnan are not.

4. Final Remarks

In developing countries like Iran, administrative records are not often available both at small and large area levels. Surveys may yield satisfactory estimates for large areas but not for small areas. Periodic censuses do not meet all demands for effective policies and planning. These limitations lead to deficiencies in official statistics. Therefore, the statistical planning activities of SCI are directed towards compensating these deficiencies by using new methods and strategies. The present study proposes a cost-effective strategy to overcome some of the limitations.

In this study, the findings support the idea that a nationwide sample design can be used instead of the separate provincial sample designs by applying suitable SAE methods. The nationwide sample design consists of nearly 13,000 persons, whereas the twenty-one separate provincial sample designs totally consist of almost 100,000 persons.

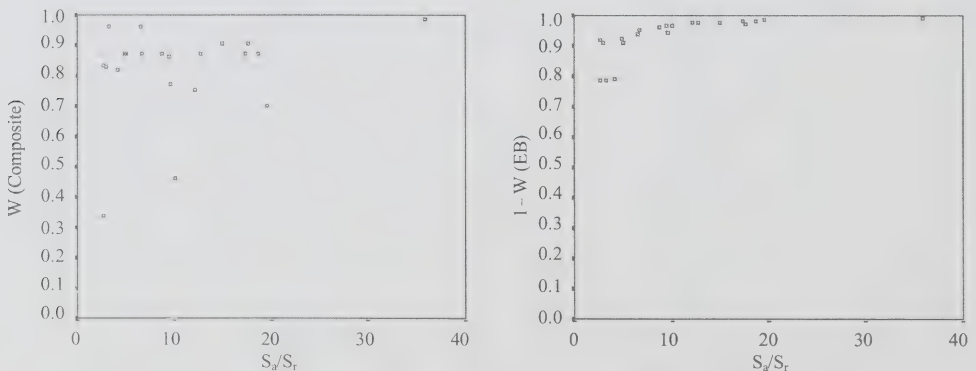


Figure 4. The weights of the direct components of composite and EB estimates against S_a/S_r .

The provincial design is the method currently used to produce provincial estimates by SCI. Using the national sample design decreases costs more than 80 percent. In addition, note that:

1. Although some SAE methods do not rely on existing sample data from all small areas, the strategy for producing provincial estimates is more appropriate when the small areas of interest are pre-specified. Therefore, the nationwide sample can be allocated to all small areas of interest to produce direct design-based estimates. It is important to adjust the sample design to accommodate the SAE methods before data collection begins. As Singh, Gambino and Mantel (1994, page 3) note

“small area needs should be recognized at the early stages of planning for large scale surveys. The sample design should include special features that enable production of reliable small area data using design or model estimators”.

Therefore, SCI must re-plan sample designs to reflect small area needs.

2. The SAE estimators usually perform well as the sample size increases. To improve provincial estimates, the nationwide sample size can be enlarged to have larger sample sizes from each province. Also, the provinces can be classified into groups with similar characteristics, such as unemployment rates, socio-demographic variables, and so on. Separate sample size would then be determined for each group.
3. Appropriate supplementary variables, which are related to the variable of interest, play a central role in improving the estimators.
 - The synthetic estimator used only one variable (age) for partitioning but it may be possible to use another variable or a combination of variables for partitioning. The post-strata in the synthetic estimator should be formed by variables that reduce variation in each post-stratum. These variables can indirectly affect the composite estimator as well.
 - The EB model can be improved with better supplementary information. Therefore it is important to try different supplementary variables to find the best model. In this work only EPA was used as the independent variable in the model, but there may be other variables that produce better estimates.
4. The composite estimator performs relatively better than the synthetic and EB estimators. However, the

results are only meant to provide a first impression of the utility of SAE methods. More research is needed to develop a generic SAE methodology in Iran. Further, the SAE methods should be applied not only in estimating unemployment rates but also in estimating other parameters, and SAE methods should be compared with the estimates coming from separate sample designs.

Acknowledgements

Research for this paper was partially supported by the Statistical Research Center of Iran. The authors are grateful for many useful and helpful comments from the referees and the Associate Editor. My heartfelt thanks should go to Jim Lepkowski for his friendly helps. The views expressed are the authors' and do not necessarily reflect those of SCI.

References

- Chand, N., and Alexander, C.H. (1995). Indirect estimation of rates and rates for small areas with continuous measurement. In *Proceeding of the Section on Survey Research Methods*, American Statistical Association, 549-554.
- Copas, J.B. (1972). Empirical Bayes methods and the repeated use of a standard. *Biometrika*, 59, 349-360.
- Cressie, N. (1989). Empirical bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.
- Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.
- Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 65-93.
- Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London.
- Gonzalez, M.F., and Hoza, C. (1978). Small area estimation with application to unemployment and housing estimation. *Journal of the American Statistical Association*, 73, 7-15.
- Gonzalez, M.F., and Waksberg, J. (1973). Estimation of the errors of synthetic estimates. Paper presented at the first meeting of the International Association of Survey Statistician, Vienna, Austria, 18-25 August.
- Levy, P.S. (1971). The use of mortality data in evaluating synthetic estimates. In *Proceedings of the American Statistical Association, Social Statistics Section*, 328-331.
- Marker, D.A. (1995). *Small area estimation: A Bayesian perspective*. Unpublished Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.
- Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.
- Pfeffermann, D. (2002). Small area estimation-New developments and directions. *International Statistical Review*, 70, 125-143.
- Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean square error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

- Purcell, N.J., and Kish, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.
- Purcell, N.J., and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, 48, 3-18.
- Rao, J.N.K. (2003a). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Rao, J.N.K. (2003b). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2, 2, 145-169.
- Schaible, W.L. (1978). Choosing weight for composite estimators for small area statistics. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 741-746.
- Schaible, W.L. (1995). Ed. *Lecture Notes in Statistics: Indirect Estimators in U.S. Federal Programs*, New York: Springer.
- Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-22.
- Thompssen, I., and Holmoy A.M.K. (1998). Combining data from surveys and administrative record system: The Norwegian experience. *International Statistical Review*, 66, 201-221.

Design Effects for Multiple Design Samples

Siegfried Gabler, Sabine Häder and Peter Lynn¹

Abstract

In some situations the sample design of a survey is rather complex, consisting of fundamentally different designs in different domains. The design effect for estimates based upon the total sample is a weighted sum of the domain-specific design effects. We derive these weights under an appropriate model and illustrate their use with data from the European Social Survey (ESS).

Key Words: Stratification; Clustering; Variance component model; Intraclass correlation coefficient; Selection probabilities.

1. Introduction

In survey research complex sample designs are often applied. These designs have features such as stratification, clustering and/or unequal inclusion probabilities, that lead to "design effects". The design effect is a measure that shows the effect of the design on the variance of an estimate. Design-based it is defined as follows (see Lohr 1999, page 239):

$$deff(plan, statistic) = \frac{V(\text{estimate from sampling plan})}{V\left(\begin{array}{c} \text{estimate from an srs with same number} \\ \text{of observation units} \end{array}\right)}$$

where srs indicates a simple random sample. The use of clustering and/or unequal inclusion probabilities typically leads to design effects greater than 1.0; in other words the variance of an estimate is increased compared to the variance of the estimate from a simple random sample with the same number of observations. The consideration of design effects is very important when deciding upon the sample size of a survey in advance. For example, if a comparative survey with different countries is planned it is very useful to have estimates of the design effects for the different countries. Then it is possible to choose the net sample sizes in a way that the precision of the estimates will be approximately equal. For this, for a certain degree of precision the sample size that would be needed under srs (effective sample size) has to be multiplied by the predicted design effect.

The European Social Survey (ESS, see www.european-socialsurvey.com) is a survey program where design effects are taken into consideration for calculating net sample sizes –aiming at the same effective sample size for each country ($n_{\text{eff}} = 1,500$). 22 countries participated in the first round of the ESS, only three of them with unclustered, equal

probability designs (srs): Denmark, Finland and Sweden. For all other countries there was the need to predict the design effect in advance of the study. For this, a model based approach (see Gabler, Häder and Lahiri 1999) can be used which distinguishes between a design effect due to unequal inclusion probabilities (term 1) and a design effect due to clustering (term 2):

$$deff = m \frac{\sum_{i=1}^I m_i w_i^2}{\left(\sum_{i=1}^I m_i w_i\right)^2} \times [1 + (b^* - 1)\rho] = deff_p \times deff_c \quad (1)$$

where m_i are respondents in the i^{th} selection probability class, each receiving a weight of w_i , ρ is the intraclass correlation coefficient and

$$b^* = \frac{\sum_{c=1}^C \left(\sum_{j=1}^{b_c} w_{cj} \right)^2}{\sum_{c=1}^C \sum_{j=1}^{b_c} w_{cj}^2}$$

where b_c is the number of observations in cluster c ($c = 1, \dots, C$) and w_{cj} is the design weight for sample element j in cluster c . (This is of course a simplification that assumes no association between y and w_i or between w_i and b^* and ignores any effects of stratification, that will tend to be beneficial and modest. See Lynn, Gabler, Häder and Laaksonen (2007, forthcoming) and Park and Lee (2004) for discussion of the sensitivity of $deff$ predictions to these assumptions; see Lynn and Gabler (2005) for discussion of alternative ways to predict $deff_c$).

In some countries the applied designs were even more complicated, consisting of fundamentally different designs in each of two independent domains. In the UK, e.g., the design was a mixture of a clustered design with unequal inclusion probabilities (in Great Britain) and an unclustered

1. Siegfried Gabler and Sabine Häder, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Postfach 12 21 55, 68072 Mannheim, Germany. E-mail: gabler@zuma-mannheim.de; Peter Lynn, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, United Kingdom. E-mail: plynn@essex.ac.uk.

sample (in Northern Ireland). In Poland, simple random samples were selected in one domain (cities and large towns), while a two-stage clustered design was applied in the second domain (all other areas). In Germany, a clustered equal-probability sample was selected in each domain (West Germany including West Berlin; East Germany), but the sampling fractions differed between the domains.

The question arose how to predict design effects for these dual design samples. As we show below, it is not simply a convex combination of the design effects for the different domains—apart from in some special cases. A general solution for multiple design samples will be presented in section 2, with illustrations of the application of this solution to prediction of design effects prior to field work (section 3) and to estimation of design effects post-field work (section 4). Section 5 concludes with discussion.

2. Design Effects for Multiple Design Samples

Let $\{C_1, \dots, C_K\}$ be a partition of the clusters into K domains. Then $Cb = \sum_{c=1}^C b_c = \sum_{k=1}^K \sum_{c \in C_k} b_c = \sum_{k=1}^K m_k = m$, where $m_k = \sum_{c \in C_k} b_c$ is the number of observations in the k^{th} domain of clusters. Let y_{cj} be the observation for sample element j in cluster c ($c=1, \dots, C$; $j=1, \dots, b_c$). The usual design-based estimator for the population mean is

$$\bar{y}_w = \frac{\sum_{c=1}^C \sum_{j=1}^{b_c} w_{cj} y_{cj}}{\sum_{c=1}^C \sum_{j=1}^{b_c} w_{cj}} = \sum_{k=1}^K \frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^C \sum_{j=1}^{b_c} w_{cj}} \bar{y}_w^{(k)}$$

where

$$\bar{y}_w^{(k)} = \frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj} y_{cj}}{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}.$$

We assume the following model M1:

$$\left. \begin{aligned} E(y_{cj}) &= \mu \\ \text{Var}(y_{cj}) &= \sigma^2 \end{aligned} \right\} \text{ for } c=1, \dots, C; j=1, \dots, b_c \quad (2)$$

$$\text{Cov}(y_{cj}, y_{c'j'}) = \begin{cases} \rho_k \sigma^2 & \text{if } c=c' \in C_k; j \neq j' \\ 0 & \text{otherwise} \end{cases} \quad k=1, \dots, K.$$

Model M1 is appropriate to account for the cluster effect with different kinds of clusters and generalises an earlier approach (see, e.g., Gabler *et al.* 1999). More general models can be found in Rao and Kleffe (1988, page 62). We define the (model) design effect as $\text{deff} = \text{Var}_{M1}(\bar{y}_w) / \text{Var}_{M2}(\bar{y})$, where $\text{Var}_{M1}(\bar{y}_w)$ is the variance of \bar{y}_w under model M1 and $\text{Var}_{M2}(\bar{y})$ is the variance of the overall

sample mean \bar{y} , defined as $\sum_{c=1}^C \sum_{j=1}^{b_c} y_{cj} / m$, computed under the following model M2:

$$\left. \begin{aligned} E(y_{cj}) &= \mu \\ \text{Var}(y_{cj}) &= \sigma^2 \end{aligned} \right\} \text{ for } c=1, \dots, C; j=1, \dots, b_c \quad (3)$$

$$\text{Cov}(y_{cj}, y_{c'j'}) = 0 \text{ for all } (c, j) \neq (c', j').$$

Note that model M2 is appropriate under simple random sampling and provides the usual expression, $\text{Var}_{M2}(\bar{y}) = \sigma^2 / m$.

Quite analogous to Gabler *et al.* (1999) we note that

$$\text{Var}_{M1} \left(\sum_{c=1}^C \sum_{j=1}^{b_c} w_{cj} y_{cj} \right) = \sigma^2 \sum_{k=1}^K \sum_{c \in C_k} \sum_{j=1}^{b_c} \left\{ w_{cj}^2 + \rho_k \sum_{j \neq j'}^{b_c} w_{cj} w_{cj'} \right\}. \quad (4)$$

Thus

$$\text{deff} = \sum_{k=1}^K \left(\frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^C \sum_{j=1}^{b_c} w_{cj}} \right)^2 \frac{m}{m_k} \text{deff}_k \quad (5)$$

where

$$\text{deff}_k = m_k \frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}^2}{\left(\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj} \right)^2} \times [1 + (b_k^* - 1) \rho_k] = \text{deff}_{pk} \times \text{deff}_{ck},$$

and

$$b_k^* = \frac{\sum_{c \in C_k} \left(\sum_{j=1}^{b_c} w_{cj} \right)^2}{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}^2}.$$

It can be seen that deff is not a convex combination of the specific $\{\text{deff}_k\}$ except in some special cases. We consider here four realistic scenarios, each representing a simplification of the general case. Only in two of these scenarios (scenarios 1 and 4) does the combination become convex:

Scenario 1: Equal weights for all units

If $w_{cj} = 1$ for all c, j , then expression (5) simplifies to:

$$\text{deff} = \sum_{k=1}^K \frac{m_k}{m} \text{deff}_k. \quad (6)$$

Scenario 2: Equal weights within each domain

If $w_{cj} = w_k$ for all $c \in C_k, j$, then expression (5) becomes:

$$deff = \sum_{k=1}^K \left(\frac{m_k w_k}{\sum_{k=1}^K m_k w_k} \right)^2 \frac{m}{m_k} deff_k. \quad (7)$$

Scenario 3: Weighted sample size proportional to domain population size

If

$$\frac{\sum_{c \in C_k} \sum_{j=1}^{b_j} w_{cj}}{\sum_{c=1}^C \sum_{j=1}^{b_c} w_{cj}} = \frac{N_k}{N},$$

where N_k is population size in domain k ; $N = \sum_{k=1}^K N_k$, then expression (5) becomes:

$$deff = \sum_{k=1}^K \left(\frac{N_k}{N} \right)^2 \frac{m}{m_k} deff_k. \quad (8)$$

Scenario 4: Unweighted sample size proportional to domain population size

If

$$\frac{m}{m_k} = \frac{N}{N_k},$$

then expression (8) becomes:

$$deff = \sum_{k=1}^K \frac{N_k}{N} deff_k. \quad (9)$$

3. Application to Prediction of Deff

In round 1 of the ESS, the sample design was a combination of two different sample designs for 5 out of 22 countries: United Kingdom, Poland, Belgium, Norway and Germany. We can apply the general formula (5) for design effects for multiple design samples to each of these cases, where $K=2$. In some cases, we can equivalently use one of the simplified expressions (6) to (9). Here we illustrate how the formulae would be used in the prediction of design effects prior to fieldwork, for the purpose of establishing the required net (respondent) sample size to achieve a prescribed precision of estimation. In each case, the approach is to predict $\{deff_k\}$ using (1) for each k and then use (5) to predict $deff$. To predict $\{deff_k\}$, the observed values of $\{w_{cj}\}$ from the ESS round 1 respondent sample are used to estimate, b^* , m_i and w_i . In other words, these could be thought of as predictions for a future survey using the same design (e.g., a future round of ESS). For illustration, we assume $\rho_k = 0.02 \forall k$ with a clustered design and $\rho_k = 0.00 \forall k$ with an unclustered design (0.02 is in fact the default value that was used for predicted design effects for clustered samples on the ESS in cases where estimates from previous surveys were not available). Our

focus here is on the application of (5). For a more detailed description of the sample designs see Häder, Gabler, Laaksonen and Lynn (2003). We use three of the ESS countries—Poland, UK and Germany—as illustrations as these designs differ between the domains in different ways. The designs of Norway and Belgium were similar to that of Poland, with equal probabilities for all units but one domain clustered and one unclustered.

3.1 Poland

In Poland, the first domain covered the population living in towns of 100,000 inhabitants or more. Within this domain, a srs of persons was selected from the population register (PESEL data base) in each region, with slight variation between regions in the sampling fraction, reflecting anticipated differences in response rate. There were 42 towns in this domain and they accounted for about 31% of the target population.

The second domain corresponded to the rest of the population—people living in towns of 99,999 inhabitants or fewer and people living in rural areas. This part of the sample was stratified and clustered (158 clusters). The sampling of this second part was based on a two-stage design: PSUs were selected with probability proportional to size. The definition of a PSU was different for urban vs. rural areas. For urban areas, a PSU was equivalent to a town, whereas for rural areas, it was equivalent to a village. In the second stage, a cluster of 12 respondents was selected in each PSU by srs.

In the first domain, $\rho_1 = 0$ and $deff_{c1} = 1$. The modest variation in selection probabilities leads to $deff_{p1} = 1.005$ and, therefore, $deff_1 = deff_{c1} \cdot deff_{p1} = 1.005$. In the second domain, the design effect due to clustering is anticipated to be $deff_{c2} = 1.18$ (based on a prediction of $b^* = 10.07$) and $deff_{p2} = 1.01$ which results in $deff_2 = deff_{c2} \cdot deff_{p2} = 1.19$. Substituting these values of $deff_k$ in (5) leads to a prediction of $deff = 1.17$.

The design for Poland differs only slightly from scenario 2 and it can be seen that in this case the simpler expression, (7), provides a reasonable prediction if we approximate the weights as follows. Domain 1 contains 37.3% of the gross sample and 31% of the target population. Thus

$$w_1 = \frac{N_1 / N}{n_1 / n} = \frac{0.310}{0.373} = 0.831$$

and

$$w_2 = \frac{N_2 / N}{n_2 / n} = \frac{0.690}{0.627} = 1.100,$$

respectively, where n_k is selected sample size in domain k ; $\sum_{k=1}^K n_k$.

Now, we can apply expression (7) to find the predicted design effect for estimates for Poland: $deff = (0.194 \cdot 1.005) + (0.821 \cdot 1.19) = 1.17$.

3.2. United Kingdom

In the UK, the ESS sample design differed between Great Britain (England, Wales, Scotland) and Northern Ireland. In Great Britain a stratified three-stage design with unequal probabilities was applied. At the first stage 162 small areas known as "postcode sectors" were selected systematically with probability proportional to the number of addresses in the sector, after implicit stratification by region and population density. At stage 2, 24 addresses were selected in each sector, leading to an equal-probability sample of addresses. At the third stage, one person aged 15+ was selected at the selected address using a Kish grid.

For Northern Ireland a simple random sample of 125 addresses was drawn from the Valuation and Land Agency's list of domestic properties. One person aged 15+ was selected at the selected address using a Kish grid. Thus, the UK sample is clustered in one domain but not in the other. In both domains, there are unequal selection probabilities.

In Great Britain we predicted $deff_{c1} = 1.20$ (based on a prediction of $b^* = 11.11$) and $deff_{p1} = 1.22$, so $deff_1 = 1.46$. In Northern Ireland we have predictions of $deff_{c2} = 1$ (by definition) and $deff_{p2} = 1.27$, so $deff_2 = 1.27$. From expression (5), $deff = 0.978 \cdot 1.46 + 0.023 \cdot 1.27 = 1.460$. It should also be noted that the selected sample sizes in the two domains were chosen to result in net sample sizes that would be approximately in proportion to the population sizes. In other words, the simplification of scenario 4 approximately holds. If we use expression (9), we get $deff = N_1 / N \cdot deff_1 + N_2 / N \cdot deff_2 = 0.97 \cdot 1.46 + 0.03 \cdot 1.27 = 1.457$, demonstrating that this provides a reasonable approximation to (5) in this case.

3.3. Germany

In Germany independent samples were selected in two domains, West Germany incl. West Berlin, and East Germany incl. East Berlin. In both domains, the sample was clustered and equal-probability, but a higher sampling fraction was used in East Germany.

At the first stage 100 communities (clusters) for West Germany, and 50 for East Germany were selected with probability proportional to the population size of the community (aged 15 years or older). The number of communities selected from each stratum was determined by a controlled rounding procedure. The number of sample points was 108 in the West, and 55 in the East (some larger communities have more than one sample point). At the second stage in each sample point there was drawn an equal number of individuals by a systematic random selection

process. This was done using the local registers of residents' registration offices.

Since the sampling design is self-weighting for both East and West Germany, but with disproportional allocation, scenario 2 applies and we can use expression (7), where

$$w_1 = w_{EAST} = \frac{N_{EAST}}{N} \frac{n}{n_{EAST}} = 0.567$$

and

$$w_2 = w_{WEST} = \frac{N_{WEST}}{N} \frac{n}{n_{WEST}} = 1.257.$$

(we note that common practice on some surveys is to scale the weights so that they sum to population sizes. This would make no difference to the application here as expression (5) involves only ratios of sums of weights).

The design effect due to clustering for each domain was predicted as $deff_{c1} = 1.39$ and $deff_{c2} = 1.35$, respectively (via predictions of $b^* = 20.56$ and 18.65 respectively), so from (7) we have

$$deff = 0.120 \cdot 1.39 + 0.991 \cdot 1.35 = 1.51.$$

It should be noted that in this case any convex combination of the domain-specific design effects will lead to a prediction of $deff$ between 1.35 and 1.39. For example, (6) would give $deff = 1.36$. This fails to take into account the differences in selection probabilities *between* the domains. With this particular design—where the *only* difference in design between domains is the difference in selection probabilities— $deff$ might alternatively be predicted by taking the convex combination and multiplying it by the prediction of $deff_p$ from the first term in expression (1), viz. $deff = 1.36 \cdot 1.09 = 1.49$. But this method is equivalent only in the special case where $\{deff_k\}$ are equal—and approximately equivalent in this case, where the variation is small.

4. Application to Estimation of $Deff$

Here we illustrate the use of expression (5) in the estimation of design effects post-fieldwork. We present estimates for 5 demographic/behavioural variables and a set of 24 attitude measures from round 1 of the European Social Survey, for the same three countries as in section 3. For comparison, we present also the estimates that would be obtained using the simpler expressions (6), (8) and (9). It can be seen that the estimates of $deff$ differ greatly between variables. This is to be expected, reflecting variation in the association of y with clusters and with selection probabilities. But here we are more interested in differences between estimation methods for the same variable.

For Germany, we see that estimators (6) and (9), which ignore variation in weights and in sampling rates between the two domains respectively, under-estimates *deff* for all variables. Estimator (8), which assumes only equal response rates in each domain, produces estimates very similar to (5). For Poland, all three simplified estimators under-estimate *deff*, though (6) perhaps performs marginally better than the other two. For UK, we observe the remarkable result that all four estimators produce almost identical estimates for every variable. The assumption in (9) (and therefore also that in (8)) holds for UK and while weights are by no means equal, the distribution of weights is very similar in each domain. It can be noted that (6) holds under a weaker assumption that

$$\frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{cj}}{\sum_{c=1}^C \sum_{j=1}^{b_c} w_{cj}} = \frac{m_k}{m},$$

i.e., that the share of the weights in each stratum equals the share of sample units. It is striking that these relationships between the estimators are consistent across all the variables considered.

5. Discussion and Conclusion

Expression (5) provides an appropriate means of combining design effects for domains with fundamentally different designs. It can be applied in estimation by estimating *deff*s in the usual way for each domain and then combining them using knowledge of the weight and domain membership of sample units. Use of (5) in the prediction of *deff*s before a survey is carried out is only slightly more demanding, requiring prediction of the share of the weights in the responding sample in each domain in addition to a method of predicting design-specific *deff*s.

Table 1
Estimates of *Deff* for Means Under 4 Estimators for 3 Countries

Estimator:	DE				GB				PL			
	(5)	(6)	(8)	(9)	(5)	(6)	(8)	(9)	(5)	(6)	(8)	(9)
<u>Demographic/behavioural</u>												
Persons in household	1.87	1.85	1.87	1.74	1.66	1.66	1.66	1.66	1.51	1.43	1.41	1.42
Years of education	3.25	2.80	3.25	2.88	2.81	2.79	2.80	2.79	1.77	1.66	1.63	1.64
Net household income	2.46	2.15	2.46	2.19	2.82	2.80	2.80	2.80	2.16	2.00	1.95	1.98
Time watching TV	2.08	1.86	2.08	1.87	2.04	2.03	2.03	2.03	1.31	1.26	1.25	1.25
Time reading newspaper	1.79	1.62	1.79	1.61	1.35	1.35	1.35	1.35	1.73	1.63	1.60	1.61
<u>Attitude measures</u>												
Discriminated by race	1.16	1.03	1.16	1.04	1.92	1.92	1.92	1.92	1.02	1.01	1.01	1.01
Discriminated by religion	1.22	1.05	1.22	1.08	1.26	1.26	1.26	1.26	1.07	1.05	1.05	1.05
General happiness	2.56	2.11	2.55	2.23	1.56	1.55	1.56	1.55	1.49	1.42	1.40	1.41
Trust in others	2.20	1.96	2.20	1.98	1.85	1.84	1.84	1.84	1.66	1.57	1.54	1.55
Trust in Euro Parliament	1.83	1.59	1.83	1.62	1.50	1.50	1.50	1.50	1.43	1.37	1.35	1.36
Trust in legal system	2.07	1.72	2.07	1.81	1.37	1.37	1.37	1.37	1.42	1.36	1.34	1.35
Trust in police	1.92	1.63	1.92	1.69	1.24	1.24	1.24	1.24	1.24	1.20	1.19	1.19
Trust in politicians	1.75	1.62	1.75	1.59	1.38	1.38	1.38	1.38	1.63	1.54	1.51	1.53
Trust in parliament	1.64	1.48	1.64	1.48	1.45	1.45	1.45	1.45	1.13	1.10	1.10	1.10
Left-right scale	1.70	1.65	1.70	1.58	1.48	1.47	1.48	1.48	1.31	1.26	1.25	1.25
Satisfaction with life	2.06	1.74	2.06	1.81	1.68	1.67	1.67	1.67	1.30	1.25	1.24	1.25
Satisfaction with education system	3.03	2.89	3.03	2.79	1.37	1.37	1.37	1.37	1.40	1.34	1.32	1.33
Satisfaction with health system	3.76	3.21	3.76	3.32	1.65	1.64	1.64	1.64	1.65	1.56	1.53	1.54
Religiosity	1.94	1.75	1.94	1.75	1.57	1.56	1.56	1.56	1.73	1.63	1.60	1.61
Attitudes to immigrants	2.77	2.68	2.77	2.57	1.92	1.92	1.92	1.92	1.89	1.76	1.73	1.74
Supports law against ethnic discrimination	2.82	2.85	2.82	2.66	1.73	1.72	1.72	1.72	2.57	2.36	2.29	2.33
Importance of family	2.17	1.99	2.17	1.97	1.19	1.19	1.19	1.19	1.21	1.17	1.17	1.17
Importance of friends	2.31	2.09	2.31	2.08	1.34	1.34	1.34	1.34	1.54	1.46	1.44	1.45
Importance of work	2.20	2.16	2.20	2.05	1.90	1.89	1.89	1.89	1.69	1.59	1.57	1.58
Support people worse off	2.70	2.47	2.70	2.45	1.35	1.35	1.35	1.35	1.78	1.67	1.64	1.66
Always obey law	2.43	2.21	2.43	2.20	1.53	1.52	1.52	1.52	2.11	1.96	1.91	1.93
Political activism	3.26	2.83	3.26	2.89	1.94	1.94	1.94	1.94	2.16	2.00	1.96	1.98
Liberalism	2.28	2.18	2.28	2.10	1.78	1.77	1.78	1.78	1.75	1.64	1.61	1.63
Participation in groups	3.75	3.04	3.75	3.24	2.26	2.25	2.25	2.25	1.82	1.71	1.68	1.69

We have shown in section 4 above that use of alternative, simpler, methods of combining the domain-specific *deffs* does not always result in good estimates. In particular, the use of a convex combination will tend to result in an under-estimation, the extent of which depends on the extent of departure from the assumptions underlying the simplified expressions. In our empirical illustration, departures were modest, but it is easy to imagine designs with greater variation between domains in mean selection probabilities or in the distribution of design weights. We would therefore recommend that estimators (6)–(9) are used only if the assumptions genuinely hold, or if the sample design data necessary to calculate (5) is not available, in which case the analyst should at least make arbitrary allowance for under-estimation based on his or her knowledge of the design.

An important issue that is outside the scope of this article is how to deal with non-response when predicting or estimating design effects for multiple design samples. The expressions throughout section 2 of this article refer to the number of observations, *i.e.*, respondent sample units, in each domain, m_k , and the calculations in sections 3 and 4 are based on predicted numbers of observations and actual numbers of observations respectively. But the natural interpretation of the differences between the four scenarios in section 2 may be in terms of sample design, where the weights are design weights. Thus, scenario 2, for example, would refer to a design that is *epsem* within domains, but where the sampling fraction is permitted to differ between domains. However, in most realistic applications non-response will occur and may well be differential both between and within domains. This is often reflected in an adjustment to the design weight. Thus, the simplification of scenario 2 would only apply if the non-response adjustment were constant within domains, in addition to the design being *epsem* within domains.

Scenario 3, if interpreted with respect to design alone, should always hold for any well-specified design in which the domains form explicit strata. Expression (8) is therefore equivalent to expression (5) in the absence of non-response. In the presence of non-response, scenario 3 requires that the (design-weighted) response rates are equal in each domain.

Similarly, scenario 4 requires that the net inclusion rate (the product of coverage rate, sampling fraction and response rate) is equal in each domain, whereas a design interpretation would not consider the response rate component.

Appropriate ways to incorporate non-response adjustment into design effect estimation and, in particular, how that might effect estimation for multiple design samples, would appear to be an area worthy of further research.

Acknowledgement

The third author is grateful to ZUMA for a guest professorship which provided the time and stimulation to write this paper and for the support of the UK Longitudinal Studies Centre at the University of Essex, which is funded by grant number H562255004 of the UK Economic and Social Research Council.

References

- Gabler, S., Häder, S. and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 105-106.
- Häder, S., Gabler, S., Laaksonen, S. and Lynn, P. (2003). The sample. Chapter 2 in *ESS 2002/2003: Technical Report*. <http://www.europeansocialsurvey.com>.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.
- Lynn, P., and Gabler, S. (2005). Approximations to b^* in the prediction of design effects due to clustering. *Survey Methodology*, 31, 101-104.
- Lynn, P., Gabler, S., Häder, S. and Laaksonen, S. (2007, forthcoming). Methods for achieving equivalence of samples in cross-national surveys. *Journal of Official Statistics*, accepted.
- Park, I., and Lee, H. (2004). Design effects for the weighted mean and total estimators under complex survey sampling. *Survey Methodology*, 30, 183-193.
- Rao, C.R., and Kleffe, J. (1988). *Estimation of Variance Components and Applications*. Amsterdam: North-Holland.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 21, No. 4, 2005

Optimal Dynamic Sample Allocation Among Strata Joseph B. Kadane	531
Evaluation of Variance Approximations and Estimators in Maximum Entropy Sampling with Unequal Probability and Fixed Sample Size Alina Matei and Yves Tillé.....	543
Implications for RDD Design from an Incentive Experiment J. Michael Brick, Jill Montaquila, Mary Collins Hagedorn, Shelley Brock Roth and Christopher Chapman	571
On the Bias in Gross Labour Flow Estimates Due to Nonresponse and Misclassification Li-Chun Zhang	591
Adjustments for Missing Data in a Swedish Vehicle Speed Survey Annica Isaksson.....	605
Conditional Ordering Using Nonparametric Expectiles Yves Aragon, Sandrine Casanova, Ray Chambers and Eve Leconte.....	617
Data Swapping as a Decision Problem Shanti Gomatam, Alan F. Karr and Ashish P. Sanil	635
An Analysis of Interviewer Effects on Screening Questions in a Computer Assisted Personal Mental Health Interview Herbert Matschinger, Sebastian Bernert and Matthias C. Angermeyer	657
Price Indexes for Elementary Aggregates: The Sampling Approach Bert M. Balk	675
Children and Adolescents as Respondents. Experiments on Question Order, Response Order, Scale Effects and the Effect of Numeric Values Associated with Response Options Marek Fuchs	701
Measuring Progress - An Australian Travelogue Jon Hall	727
Quality on Its Way to Maturity: Results of the European Conference on Quality and Methodology in Official Statistics (Q2004) Werner Grünewald and Thomas Körner	747
Editorial Collaborators.....	761

Volume 33, No. 4, December/décembre 2005

Serge TARDIF, François BELLAVANCE and Constance VAN EEDEN A nonparametric procedure for the analysis of balanced crossover designs.....	471
José E. CHACÓN and Alberto RODRÍGUEZ-CASAL On the L_1 -consistency of wavelet density estimates.....	489
Rohana J. KARUNAMUNI and Tom ALBERTS A generalized reflection method of boundary correction in kernel density estimation.....	497
Ana M. BIANCO, Marta Garcia BEN and Víctor J. YOHAI Robust estimation for linear regression with asymmetric errors.....	511
Xin GAO and Mayer ALVO A nonparametric test for interaction in two-way layouts.....	529
Lan WANG and Xiao-Hua ZHOU A fully nonparametric diagnostic test for homogeneity of variances.....	545
Guosheng YIN and Joseph G. IBRAHIM Cure rate models: a unified approach.....	559
George ILIOPOULOS, Dimitris KARLIS and Ioannis NTZOUFRAS Bayesian estimation in Kibble's bivariate gamma distribution.....	571
Dongchu SUN and Paul L. SPECKMAN A note on nonexistence of posterior moments.....	591
Forthcoming papers/Articles à paraître.....	609

Volume 34, No. 1, March/mars 2006

Angelo J. CANTY, Anthony C. DAVISON, David V. HINKLEY and Valérie VENTURA Bootstrap diagnostics and remedies.....	5
Christian LÉGER and Brenda MACGIBBON On the bootstrap in cube root asymptotics.....	29
Min TSAO and Changbao WU Empirical likelihood inference for a common mean in the presence of heteroscedasticity.....	45
Ricardo CAO and Ingrid VAN KEILEGOM Empirical likelihood tests for two-sample problems via nonparametric density estimation.....	61
Jinhong YOU, Gemain CHEN and Yong ZHOU Une vraisemblance empirique par bloc pour les modèles de régression partiellement linéaires longitudinaux.....	79
Xuewen LU, Gemai CHEN, Radhey S. SINGH and Peter X.-K. SONG A class of partially linear single-index survival models.....	97
Michael J. EVANS, Irwin GUTTMAN and Tim SWARTZ Optimality and computations for relative surprise inferences.....	113
Abdelouahab BIBI and Antony GAUTIER L_2 -properties and estimation of purely bilinear and strictly superdiagonal time series models with periodic coefficients.....	131
Wenceslao GONZÁLEZ-MANTEIGA and Ana PÉREZ-GONZÁLEZ Goodness-of-fit tests for linear regression models with missing response data.....	149
Jae Kwang KIM and Hyeonah PARK Imputation using response probability.....	171
Forthcoming papers/Articles à paraître.....	183

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, N° 1) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préféablement Word. Une version papier pourrait être requise pour les formules et graphiques.

- 1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

- 2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

- 3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0, I, 1).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.

- 4. **Figures et tableaux**
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

- 5. **Bibliographie**
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).
 - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Volume 33, No. 4, December/décembre 2005

Serge TARDIF, François BELLAVANCE and Constance VAN EEDEN	471
A nonparametric procedure for the analysis of balanced crossover designs	
José E. CHACÓN and Alberto RODRÍGUEZ-CASAL	489
On the L_1 -consistency of wavelet density estimates	
Rohana J. KARUNAMUNI and Tom ALBERTS	497
A generalized reflection method of boundary correction in kernel density estimation	
Ana M. BIANCO, Maria Garcia BEN and Victor J. YOHAI	511
Robust estimation for linear regression with asymmetric errors	
Xin GAO and Mayer ALVO	529
A nonparametric test for interaction in two-way layouts	
Lan WANG and Xiao-Hua ZHOU	545
A fully nonparametric diagnostic test for homogeneity of variances	
Guosheng YIN and Joseph G. IBRAHIM	559
Cure rate models: a unified approach	
George LIPOULOS, Dimitris KARLIS and Ioannis NTZOUFRAS	571
Bayesian estimation in Kibble's bivariate gamma distribution	
Dongchu SUN and Paul L. SPECKMAN	591
A note on nonexistence of posterior moments	
Forthcoming papers/Articles à paraître	609

Volume 34, No. 1, March/mars 2006

Angelo J. CANTY, Anthony C. DAVISON, David V. HINKLEY and Valérie VENTURA	5
Bootstrap diagnostics and remedies	
Christian LÉGER and Brenda MACGIBBON	29
On the bootstrap in cube root asymptotics	
Min TSAO and Changbao WU	45
Empirical likelihood inference for a common mean in the presence of heteroscedasticity	
Ricardo CAO and Ingrid VAN KEILEGOM	61
Empirical likelihood tests for two-sample problems via nonparametric density estimation	
Jinhong YU, Gemai CHEN and Yong ZHOU	79
Une vraisemblance empirique par bloc pour les modèles de régression partiellement linéaires longitudinaux	
Xuewen LU, Gemai CHEN, Radhey S. SINGH and Peter X.-K. SONG	97
A class of partially linear single-index survival models	
Michael J. EVANS, Irwin GUTTMAN and Tim SWARTZ	113
Optimality and computations for relative surprise inferences	
Abdelouahab BIBI and Antony GAUTIER	131
L_2 -properties and estimation of purely bilinear and strictly superdiagonal time series models with periodic coefficients	
Wenceslao GONZÁLEZ-MANTEIGA and Ana PEREZ-GONZÁLEZ	149
Goodness-of-fit tests for linear regression models with missing response data	
Jae Kwang KIM and Hyeonah PARK	171
Imputation using response probability	
Forthcoming papers/Articles à paraître	183

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOs is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 21, No. 4, 2005

Optimal Dynamic Sample Allocation Among Strata	531
Joseph B. Kadane	
Evaluation of Variance Approximations and Estimators in Maximum Entropy Sampling with Unequal Probability and Fixed Sample Size	543
Alina Matei and Yves Tillé	
Implications for RDD Design from an Incentive Experiment	571
J. Michael Brick, Jill Montaguila, Mary Collins Hagedorn, Shelley Brock Roth and Christopher Chapman	
On the Bias in Gross Labour Flow Estimates Due to Nonresponse and Misclassification	591
Li-Chun Zhang	
Adjustments for Missing Data in a Swedish Vehicle Speed Survey	605
Annica Isaksson	
Conditional Ordering Using Nonparametric Expectiles	617
Yves Aragon, Sandrine Casanova, Ray Chambers and Eve Lecomte	
Data Swapping as a Decision Problem	635
Shanti Gomataam, Alan F. Karr and Ashish P. Sanil	
An Analysis of Interviewer Effects on Screening Questions in a Computer Assisted Personal Health Interview	657
Herbert Matschinger, Sebastian Bernert and Mathias C. Angermeyer	
Price Indexes for Elementary Aggregates: The Sampling Approach	675
Bert M. Balk	
Children and Adolescents as Respondents. Experiments on Question Order, Response Order, Scale Effects and the Effect of Numeric Values Associated with Response Options	701
Marek Fuchs	
Measuring Progress - An Australian Travelogue	727
Jon Hall	
Quality on Its Way to Maturity: Results of the European Conference on Quality and Methodology in Official Statistics (Q2004)	747
Werner Grünewald and Thomas Körner	
Editorial Collaborators	761

All inquiries about submissions and subscriptions should be directed to jos@scb.se

Bibliographie

Gabler, S., Häder, S. et Lahiri, P. (1999). Justification à base de modèle de la formule de Kish pour les effets de plan de sondage liés à la pondération et à l'effet de grappe. *Techniques d'enquête*, 25, 119-120.

Häder, S., Gabler, S., Laaksonen, S. et Lynn, P. (2003). The sample, Chapitre 2 dans *ESS 2002/2003: Rapport technique*. <http://www.europeansocialsurvey.com>.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.

Lynn, P., et Gabler, S. (2005). Approximations de b^* dans la prévision des effets du plan dus à la mise en grappes. *Techniques d'enquête*, 31, 109-113.

Lynn, P., Gabler, S., Häder, S. et Laaksonen, S. (2007, à paraître). Methods for achieving equivalence of samples in cross-national surveys. *Journal of Official Statistics*, accepté.

Park, I., et Lee, H. (2004). Effets de plan pour les estimateurs pondérés de la moyenne et du total sous échantillonnage complexe. *Techniques d'enquête*, 30, 205-216.

Kao, C.R., et Kleffe, J. (1988). *Estimation of Variance Components and Applications*. Amsterdam: North-Holland.

Remerciements

Le troisième auteur remercie la ZUMA pour le poste de professeur invité qui lui a permis de trouver le temps et les conditions propices pour rédiger le présent article et remercie aussi de son soutien le UK Longitudinal Studies Centre de l'Université d'Essex, qui est financé par la subvention H562255004 de l'Economic and Social Research Council du Royaume-Uni.

plan et, en particulier, l'effet que cette correction pourrait avoir sur l'estimation dans le cas d'échantillons à plans multiples, semble être un domaine qui mérite d'être exploré lors de futures études.

Tableau 1
Estimations de *Deff* pour les moyennes sous quatre estimateurs pour trois pays

Estimateur :	Allemagne			Royaume-Uni			Pologne		
	(5)	(6)	(8)	(5)	(6)	(8)	(5)	(6)	(9)
Nbre d'années d'études	3,25	2,80	3,25	2,81	2,79	2,80	2,79	1,77	1,63
Revenu net du ménage	2,46	2,15	2,46	2,19	2,82	2,80	2,80	2,16	1,95
Temps passé à regarder la TV	2,08	1,86	2,08	1,87	2,04	2,03	2,03	1,31	1,26
Temps passé à lire le journal	1,79	1,62	1,79	1,61	1,35	1,35	1,35	1,73	1,63
Mesures d'attitude	1,16	1,03	1,16	1,04	1,92	1,92	1,92	1,02	1,01
Discrimination selon la race	1,22	1,05	1,22	1,08	1,26	1,26	1,26	1,07	1,05
Discrimination selon la religion	2,11	2,55	2,23	1,56	1,55	1,55	1,55	1,49	1,47
Etat de bonheur général	2,20	1,96	2,20	1,98	1,85	1,84	1,84	1,66	1,57
Confiance dans les autres	2,20	1,96	2,20	1,98	1,85	1,84	1,84	1,66	1,54
Confiance dans le Parlement européen	1,83	1,59	1,83	1,62	1,50	1,50	1,50	1,43	1,37
Confiance dans le système juridique	2,07	1,72	2,07	1,81	1,37	1,37	1,37	1,42	1,36
Confiance dans la police	1,92	1,63	1,92	1,69	1,24	1,24	1,24	1,24	1,20
Confiance dans les politiciens	1,75	1,62	1,75	1,59	1,38	1,38	1,38	1,63	1,54
Confiance dans le Parlement	1,64	1,48	1,64	1,48	1,45	1,45	1,45	1,13	1,10
Échelle gauche-droite	1,70	1,65	1,70	1,58	1,48	1,47	1,48	1,31	1,26
Satistfaction à l'égard de la vie	2,06	1,74	2,06	1,81	1,68	1,67	1,67	1,30	1,25
Satistfaction à l'égard du système d'éducation	3,03	2,89	3,03	2,79	1,37	1,37	1,37	1,40	1,32
Satistfaction à l'égard du système de santé	3,76	3,21	3,76	3,32	1,65	1,64	1,64	1,65	1,53
Attitude religieuse	1,94	1,75	1,94	1,75	1,57	1,56	1,56	1,73	1,63
Attitude à l'égard des immigrants	2,77	2,68	2,82	2,57	1,92	1,92	1,92	1,89	1,76
Appuie une loi contre la discrimination ethnique	2,82	2,85	2,82	2,66	1,73	1,72	1,72	2,57	2,36
Importance de la famille	2,17	1,99	2,17	1,97	1,19	1,19	1,19	1,21	1,17
Importance des amis	2,31	2,09	2,31	2,08	1,34	1,34	1,34	1,54	1,44
Importance du travail	2,20	2,16	2,20	2,05	1,90	1,89	1,89	1,69	1,59
Aide les personnes moins bien nantes	2,70	2,47	2,70	2,45	1,35	1,35	1,35	1,78	1,67
Respecte toujours la loi	2,43	2,21	2,43	2,20	1,52	1,52	1,52	2,11	1,96
Activisme politique	2,28	2,18	2,28	2,10	1,78	1,77	1,78	1,75	1,64
Libéralisme	3,75	3,04	3,75	3,24	2,26	2,25	2,25	1,82	1,71
Participation à des groupes	3,75	3,04	3,75	3,24	2,26	2,25	2,25	1,82	1,71

de sondage. Donc, la simplification du scénario 2 ne serait applicable que si l'ajustement pour la non-réponse était constant dans les domaines, outre la sélection avec *probabilités égales* dans les domaines.

Le scénario 3, s'il est interprété uniquement par rapport au plan de sondage, devrait être vérifié pour tout plan de sondage bien spécifié dans lequel les domaines forment des strates explicites. L'expression (8) est par conséquent équivalente à l'expression (5), en l'absence de non-réponse. En présence de non-réponse, le scénario 3 exige que les taux de réponse (pondérés par les poids de sondage) soient égaux dans chaque domaine. De même, le scénario 4 demande que le taux d'inclusion net (produit du taux de couverture, de la fraction d'échantillonnage et du taux de réponse) soit le même dans chaque domaine, tandis qu'une interprétation basée sur le plan de sondage ne tiendrait pas compte de la composante du taux de réponse.

La recherche de moyens appropriés d'intégrer l'ajustement pour la non-réponse dans l'estimation de l'effet de

Une question importante qui dépasse le cadre du présent article est celle de savoir comment traiter la non-réponse lors de la prédiction ou de l'estimation des effets de plan pour les échantillons à plans multiples. Les expressions présentées à la section 2 se rapportent au nombre d'observations (unités d'échantillonnage répondantes) dans chaque domaine, m_i , et les calculs présentés aux sections 3 et 4 sont fondés sur les nombres prévus et réels d'observations, respectivement. Cependant, l'interprétation naturelle des différences entre les quatre scénarios de la section 2 pourrait se faire en fonction du plan de sondage, où les pondérations sont les poids de sondage. Donc, le scénario 2, par exemple, aurait trait à un plan de sondage avec *probabilités de sélection égales* dans les domaines, mais où la fraction d'échantillonnage peut varier selon le domaine. Cependant, dans la plupart des applications réelles, il y aura une non-réponse qui pourrait fort bien différer entre les domaines, ainsi que dans les domaines, situation qui est souvent reflétée par un ajustement du poids

(Nous notons qu'il est courant, dans certaines enquêtes, de rééchantillonner les pondérations afin que leur somme soit égale à la taille de population. Cette pratique n'aurait aucune incidence ici, car l'expression (5) ne comprend que des ratios de sommes de pondérations).

Pour chaque domaine, nous avons prédit les effets de plan $deff_{ci} = 1,39$ et $deff_{c2} = 1,35$, respectivement (d'après les prédictions que $b^* = 20,56$ et 18,65, respectivement), si bien qu'il découle de (7) que

$$deff = 0,120 \cdot 1,39 + 0,991 \cdot 1,35 = 1,51.$$

Il convient de souligner que dans ce cas, toute combinaison convexe des effets de plan de domaine produira une prédiction de $deff$ comprise entre 1,35 et 1,39. Par exemple, (6) donnerait $deff = 1,36$. Ce résultat ne tient pas compte des différences de probabilité de sélection *entre* les domaines. Dans le cas du plan de sondage examiné ici, où la *seule* différence de plan de sondage entre les domaines est la différence de probabilité de sélection, $deff$ pourrait aussi être prédit en prenant la combinaison convexe et en la multipliant par la prédiction de $deff^p$ pour le premier terme de l'expression (1), c'est-à-dire $deff = 1,36 \cdot 1,09 = 1,49$. Cette méthode n'est toutefois équivalente que dans le cas particulier où les $\{deff_k\}$ sont égaux, et approximativement équivalente ici, où la variation est faible.

4. Application à l'estimation de $Deff$

Nous allons maintenant illustrer l'utilisation de l'expression (5) pour estimer les effets de plan après le travail sur le terrain. Nous présentons des estimations pour cinq variables démographiques/de comportement et un ensemble de 24 mesures d'attitude provenant du premier cycle de l'Enquête sociale européenne (ESS) pour les trois mêmes pays qu'à la section 3. Aux fins de comparaison, nous présentons aussi les estimations que l'on obtiendrait en utilisant les expressions plus simples (6), (8) et (9). Les résultats montrent que les estimations de $deff$ diffèrent considérablement selon la variable. Cette situation, qui est prévisible, reflète la variation de l'association de y avec les groupes et les probabilités de sélection. Mais ici, nous nous intéressons principalement aux différences entre les méthodes d'estimation pour une même variable.

Dans le cas de l'Allemagne, nous voyons que les estimateurs (6) et (9), qui ne tiennent pas compte de la variation de la pondération et des taux d'échantillonnage entre les deux domaines, sous-estiment $deff$ pour toutes les variables. L'estimateur (8), qui repose uniquement sur l'hypothèse que les taux de réponse sont égaux dans chaque domaine, produit des estimations fort semblables à (5). Pour la Pologne, les trois estimateurs simplifiés sous-estiment $deff$, quoique (6) pourrait produire des résultats légèrement

5. Discussion et conclusion

c'est-à-dire que la part des pondérations dans chaque strate est égale à la part des unités d'échantillonnage. Il est trappant que ces relations entre les estimateurs soient convergentes pour l'ensemble des variables considérées.

$$\frac{\sum_{b^*} \sum_{j=1}^{c-1} w_{cj}}{\sum_{b^*} w_{cj}} = \frac{\sum_{c \in C_i} \sum_{j=1}^{c-1} w_{cj}}{\sum_{b^*} w_{cj}},$$

meilleurs que les deux autres. Pour ce qui est du Royaume-Uni, nous obtenons le résultat remarquable que les quatre estimateurs produisent des estimations presque identiques pour chaque variable. L'hypothèse qui sous-tend (9) (et par conséquent également celle qui sous-tend (8)) tient pour le Royaume-Uni et, bien que les pondérations soient loin d'être égales, leur distribution est fort semblable dans chaque domaine. Il convient de souligner que (6) est vérifiée sous une hypothèse plus faible que

L'expression (5) offre un moyen approprié de combiner les effets de plan pour des domaines pour lesquels les plans de sondage sont fondamentalement différents. Elle peut être appliquée en estimant l'effet de plan $deff$ de la manière habituelle pour chaque domaine, puis en les combinant d'après les données sur la pondération et l'appartenance des unités d'échantillonnage au domaine. L'utilisation de (5) pour prédire les effets de plan $deff$ avant qu'une enquête soit réalisée est à peine plus difficile. Elle nécessite la prédiction de la part des pondérations dans l'échantillon de répondants dans chaque domaine, en plus d'une méthode de prédiction des $deff$ propres aux divers plans de sondage.

Nous avons montré à la section 4 que précède que l'utilisation d'autres méthodes, plus simples, de combinaisons des $deff$ de domaine ne produit pas toujours de bonnes estimations. Plus précisément, l'utilisation d'une combinaison convexe aura tendance à causer une sous-estimation dont l'importance dépend de l'écart par rapport aux hypothèses qui sous-tendent les expressions simplifiées. Dans notre exemple empirique, les écarts étaient modérés, mais il est facile d'imaginer des plans de sondage où les variations des probabilités de sélection moyennes ou de la distribution des poids de sondage selon le domaine sont plus importantes. Nous devrions par conséquent recommander de n'utiliser les estimateurs (6) à (9) que si les hypothèses sont vraiment vérifiées ou que les données sur le plan de sondage nécessaires pour calculer (5) ne sont pas disponibles, auquel cas l'analyste devrait au moins tenir compte arbitrairement d'une sous-estimation en se basant sur sa connaissance du plan de sondage.

les secondes, il s'agissait d'un village. À la deuxième phase, une grappe de 12 répondants a été sélectionnée par cas dans

chaque UPE.

Dans le premier domaine, $p_1 = 0$ et $def_{c1} = 1$. La

variation modérée des probabilités de sélection mène à $def_{p1} = 1,005$ et, par conséquent, $def_{f1} = def_{c1} \cdot def_{p1} = 1,005$. Dans le deuxième domaine, l'effet de plan dû à la mise en grappes prévu est $def_{c2} = 1,18$ (d'après la prédiction que $b^* = 10,07$) et $def_{p2} = 1,01$, ce qui donne $def_{f2} = def_{c2} \cdot def_{p2} = 1,19$. La substitution de ces valeurs à def_{f1} dans (5) donne un effet de plan prévu égal à $def_{f1} = 1,17$.

Le plan de sondage appliqué en Pologne ne diffère que légèrement du scénario 2 et, dans ce cas, nous voyons que l'expression plus simple, (7), donne une prédiction raisonnable si nous calculons la valeur approximative des pondérations comme suit. Le domaine 1 contient 37,3 % de l'échantillon brut et 31 % de la population cible. Donc

$$w_1 = \frac{N_1/N}{n_1/n} = \frac{0,310}{0,373} = 0,831$$

et

$$w_2 = \frac{N_2/N}{n_2/n} = \frac{0,690}{0,627} = 1,100,$$

respectivement, où n_k est la taille de l'échantillon sélectionné dans le domaine k ; $\sum_{k=1}^K n_k$.

Maintenant, nous pouvons appliquer l'expression (7) pour calculer l'effet de plan prévu pour les estimations pour la Pologne : $def = (0,194 \cdot 1,005) + (0,821 \cdot 1,19) = 1,17$.

3.2. Royaume-Uni

Au Royaume-Uni, le plan de sondage de l'ESS n'était pas le même pour la Grande-Bretagne (Angleterre, Pays de Galles, Ecosse) que pour l'Irlande du Nord. En Grande-Bretagne, on a utilisé un plan stratifié à trois degrés avec probabilités de sélection inégales. À la première étape, 162 petits secteurs, appelés « secteurs de code postal » ont été sélectionnés systématiquement avec probabilité proportionnelle au nombre d'adresses dans le secteur, après stratification implicite selon la région et la densité de population. À la deuxième étape, 24 adresses ont été sélectionnées dans chaque secteur, ce qui a produit un échantillon avec probabilités égales d'adresses. À la troisième étape, une personne de 15 ans ou plus a été sélectionnée à chaque adresse échantillonnée au moyen d'une grille de Kish.

Pour l'Irlande du Nord, un échantillon aléatoire simple de 125 adresses a été tiré d'après la liste de propriétés privées de l'Agence d'évaluation foncière (Valuation and Land Agency). Puis, une personne de 15 ans ou plus a été sélectionnée à chaque adresse échantillonnée en utilisant une grille de Kish. Donc, l'échantillon du Royaume-Uni est

3.3. Allemagne

En Allemagne, des échantillons indépendants ont été sélectionnés dans deux domaines, c'est-à-dire l'Allemagne de l'Ouest, y compris Berlin Ouest et l'Allemagne de l'Est, y compris Berlin Est. Dans chaque domaine, on a sélectionné un échantillon par grappes avec probabilités égales, mais en utilisant une plus grande fraction d'échantillonnage pour l'Allemagne de l'Est.

À la première étape, 100 communautés (grappes) ont été sélectionnées pour l'Allemagne de l'Ouest et 50 pour l'Allemagne de l'Est avec probabilité proportionnelle à la taille de la population de la communauté (personnes de 15 ans et plus). Le nombre de communautés sélectionnées dans chaque strate a été déterminé par une méthode d'arrondissement contrôlé. Le nombre de points d'échantillonnage était de 108 à l'Ouest et de 55 à l'Est (pour certaines grandes collectivités, on a utilisé plus d'un point d'échantillonnage). À la deuxième étape, pour chaque point d'échantillonnage, on a sélectionné un nombre égal de personnes par échantillonnage aléatoire systématique, d'après les registres locaux des bureaux d'enregistrement des résidents.

Puisque le plan de sondage est autopondéré aussi bien pour l'Allemagne de l'Est que de l'Ouest, mais que la répartition est disproportionnelle, nous pouvons appliquer le scénario 2 et utiliser l'expression (7), où

$$w_1 = w_{Est} = \frac{N_{Est}}{n_{Est}} = \frac{0,567}{n}$$

et

$$w_2 = w_{Ouest} = \frac{N_{Ouest}}{n_{Ouest}} = \frac{1,257}{n}$$

On peut voir que $deff_k$ n'est pas une combinaison convexe des effets de plan spécifiques $\{deff_k\}$, sauf dans des cas particuliers. Nous considérons ici quatre scénarios raisonnables, chacun représentant une simplification du cas général. La combinaison ne devient convexe que dans deux de ces scénarios (1 et 4) :

Scénario 1 : Même pondération pour toutes les unités

Si $w_{cj} = 1$ pour tous c, j , alors l'expression (5) se simplifie comme suit :

$$(6) \quad deff = \sum_{k=1}^K \frac{m}{m_k} deff_k.$$

Scénario 2 : Même pondération des unités dans chaque domaine

Si $w_{cj} = w_k$ pour tous $c \in C_k, j$, alors l'expression (5) devient :

$$(7) \quad deff = \sum_{k=1}^K \left(\frac{\sum_{c \in C_k} \sum_{j=1}^{w_{cj}} m_k w_k}{m_k w_k} \right) deff_k.$$

Scénario 3 : Taille d'échantillon pondéré proportionnelle à la taille de la population du domaine

Si

$$\frac{\sum_{c \in C_k} \sum_{j=1}^{w_{cj}} w_{cj}}{N_k} = \frac{\sum_{b \in B_k} w_{bj}}{N_k},$$

où N_k est la taille de population dans le domaine k ; $N = \sum_{k=1}^K N_k$, alors l'expression (5) devient :

$$(8) \quad deff = \sum_{k=1}^K \left(\frac{N}{N_k} \right) \frac{m}{m_k} deff_k.$$

Scénario 4 : Taille d'échantillon non pondéré proportionnelle à la taille de population du domaine

Si

$$\frac{m}{N} = \frac{m_k}{N_k},$$

alors l'expression (8) devient :

$$(9) \quad deff = \sum_{k=1}^K \frac{N}{N_k} deff_k.$$

3. Application à la prédiction de $Deff$

Lors du premier cycle de l'ESS, le plan de sondage était une combinaison de deux plans différents pour 5 des 22 pays, à savoir le Royaume-Uni, la Pologne, la Belgique, la Norvège et l'Allemagne. Nous pouvons appliquer la

formule générale (5) des effets de plan pour échantillon à plans multiples à chacun de ces cas, où $K=2$. Pour certains d'entre eux, nous pouvons utiliser indirectement l'un des expressions simplifiées (6) à (9). Ici, nous illustrons comment la formule serait utilisée pour prédire les effets de plan avant le travail sur le terrain en vue d'établir la taille d'échantillon nette (répondants) requise pour atteindre une précision d'estimation préalable. Dans chaque cas, l'approche consiste à prédire $\{deff_k\}$ en utilisant (1) pour chaque k , puis à utiliser (5) pour prédire $deff$. Afin de prédire $\{deff_k\}$, nous utilisons les valeurs observées de $\{w_{bj}\}$ provenant de l'échantillon de répondants du premier cycle de l'ESS pour estimer b^*, m^* et w^* . En d'autres termes, nous pourrions considérer que ces prédictions sont faites pour une future enquête basée sur le même plan de sondage (par exemple, un futur cycle de l'ESS). À titre d'exemple, nous supposons que $p_k = 0,02 \forall k$ avec un plan de sondage par grappes et $p_k = 0,004 \forall k$ avec un plan sans mise en grappes (0,02 dans l'ESS dans les cas où l'on ne disposait pas d'estimations d'après des enquêtes antérieures). Ici, nous nous concentrons sur l'application de l'équation (5). Pour une description plus détaillée des plans de sondage, voir Häder, Gabler, Laaksonen et Lynn (2003). Nous choisissons comme exemples trois des pays participants à l'ESS, la Pologne, le Royaume-Uni et l'Allemagne, car ils ont utilisé des plans de sondage multiples dont les différences entre domaines ne sont pas les mêmes. Les plans de sondage utilisés par la Norvège et la Belgique étaient semblables à celui de la Pologne, avec probabilités d'inclusion égales pour toutes les unités, mais mise en grappes dans un domaine et non dans l'autre.

3.1 Pologne

En Pologne, le premier domaine couvrait la population des villes de 100 000 habitants et plus. Dans ce domaine, des personnes ont été sélectionnées par EAS d'après le registre de population (base de données PSESL) dans chaque région, avec application d'une fraction d'échantillonage légèrement différente selon la région afin de refléter les différences attendues de taux de réponse. Ce domaine comprenait 42 villes qui représentaient environ 31 % de la population cible.

Le deuxième domaine correspondait au reste de la population, c'est-à-dire les personnes vivant dans les villes de 99 999 habitants et moins et dans les régions rurales. Cette partie de l'échantillon a été stratifiée et mise en grappes (158 grappes). L'échantillonage de cette deuxième partie a été réalisé selon un plan à deux degrés où les UPE ont été sélectionnées avec probabilité proportionnelle à la taille. La définition de l'UPE n'était pas la même pour les régions urbaines que pour les régions rurales. Pour les premières, l'UPE correspondait à une ville, tandis que pour

ces hypothèses; voir Lym et Gabler (2005) pour une discussion de divers moyens de prédire $deff_k$.

Dans certains pays, les plans de sondage utilisés étaient encore plus compliqués, comprenant des plans fondamentallement différents dans chacun des deux domaines indépendants. Au Royaume-Uni, par exemple, il s'agissait du mélange d'un plan par grappes avec probabilités d'inclusion inégales (en Grande-Bretagne) et d'un échantillon sans mise en grappes (en Irlande du Nord). En Pologne, des échantillons aléatoires simples ont été sélectionnés dans un domaine (villes grandes et moyennes), tandis qu'un plan à deux degrés avec mise en grappes a été appliqué au deuxième domaine (toutes les autres régions). En Allemagne, un échantillon par grappes avec probabilités de sélection égales a été sélectionné dans chaque domaine (Allemagne de l'Ouest, y compris Berlin Ouest; Allemagne de l'Est), mais les fractions d'échantillonnage n'étaient pas les mêmes dans les deux domaines.

La question de la prédiction des effets de plan s'est donc posée pour ces échantillons à plan de sondage double. Comme nous le montrons plus loin, il ne s'agit pas simplement d'une combinaison convexe des effets de plan pour les divers domaines, à part dans des cas spéciaux. Nous présentons une solution générale pour les échantillons à plans de sondage multiples à la section 2, ainsi que des exemples d'application de cette solution afin de prédire les effets de plan avant le travail sur le terrain (section 3) et après le travail sur le terrain (section 4). À la section 5, nous concluons par une discussion.

2. Effets de plan pour les échantillons à plans de sondage multiples

Soit $\{C_1, \dots, C_K\}$ une partition des grappes en K domaines. Alors, $C_j^c = \sum_{k=1}^{c-1} C_k$, $b_c^c = \sum_{k=1}^{c-1} b_k = m$, où $m_k = \sum_{c \in C_k} b_c$ est le nombre d'observations dans le k^e domaine de grappes. Soit $y_{c,j}$ l'observation pour l'unité d'échantillonnage j dans la grappe c ($c = 1, \dots, C$; $j = 1, \dots, b_c$). L'estimateur habituel fondé sur le plan de sondage de la moyenne de population est

$$\bar{y}^{(K)} = \frac{\sum_{c \in C_K} \sum_{j=1}^{b_c} y_{c,j}}{\sum_{c \in C_K} b_c} = \frac{\sum_{c \in C_K} \sum_{j=1}^{b_c} w_{c,j}^{(K)}}{\sum_{c \in C_K} w_c^{(K)}}.$$

où

$$\bar{y}^{(K)} = \frac{\sum_{c=1}^K \sum_{j=1}^{b_c} w_{c,j}^{(K)}}{\sum_{c=1}^K \sum_{j=1}^{b_c} w_c^{(K)}} = \frac{\sum_{c=1}^K \sum_{j=1}^{b_c} w_{c,j}^{(K)}}{\sum_{c=1}^K \sum_{j=1}^{b_c} w_c^{(K)}}.$$

où

$$deff_k = m_k \frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_c^{(K)}}{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{c,j}^{(K)}} \times [1 + (b_k^* - 1) \rho_k] = deff_{pk} \times deff_{ck},$$

et

$$b_k^* = \frac{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_c^{(K)}}{\sum_{c \in C_k} \sum_{j=1}^{b_c} w_{c,j}^{(K)}}.$$

Donc

$$\text{Var}_{M1} \left(\sum_{c=1}^K \sum_{j=1}^{b_c} w_{c,j}^{(K)} y_{c,j}^{(K)} \right) = \sigma^2 \sum_{c=1}^K \sum_{j=1}^{b_c} w_c^{(K)} + \rho_K \sum_{c=1}^K \sum_{j=1}^{b_c} w_{c,j}^{(K)} \left\{ E(y_{c,j}^{(K)}) = \mu \right\} \quad (4)$$

notons que De façon fort semblable à Gabler et coll. (1999), nous habituelle $\text{Var}_{M2}(\bar{y}) = \sigma^2 / m$.

Souignons que le modèle M2 est approprié sous échantillonnage aléatoire simple et donne l'expression

$$\text{Cov}(y_{c,j}^{(K)}, y_{c',j'}^{(K)}) = 0 \quad \text{pour tous les } (c, j) \neq (c', j'). \quad (3)$$

Le modèle M1 convient pour tenir compte de l'effet de grappe avec divers types de grappes et généralise une approche antérieure (voir, par exemple, Gabler et coll. 1999). Des modèles plus généraux sont décrits dans Rao et Kieffe (1988, pages 62). Nous définissons l'effet de plan (par rapport au modèle) comme étant $deff = \text{Var}_{M1}(\bar{y}^{(K)}) / \text{Var}_{M2}(\bar{y})$, où $\text{Var}_{M1}(\bar{y}^{(K)})$ est la variance de $\bar{y}^{(K)}$ sous le modèle M1 et $\text{Var}_{M2}(\bar{y})$ est la variance de la moyenne globale d'échantillon \bar{y} , définie comme étant $\sum_{c=1}^K \sum_{j=1}^{b_c} w_{c,j}^{(K)}$.

Nous postulons le modèle M1 suivant :

$$\text{Cov}(y_{c,j}^{(K)}, y_{c',j'}^{(K)}) = \begin{cases} \rho_K \sigma^2 & \text{si } c' = c \text{ et } j' \neq j \\ 0 & \text{autrement} \end{cases} \quad (2)$$

$$\left\{ \begin{array}{l} E(y_{c,j}^{(K)}) = \mu \\ \text{Var}(y_{c,j}^{(K)}) = \sigma^2 \end{array} \right\} \quad \text{pour } c = 1, \dots, C; j = 1, \dots, b_c$$

Effets de plan pour les échantillons à plans de sondage multiples

Siegfried Gabler, Sabine Häder et Peter Lynn

Résumé

Dans certaines situations, le plan de sondage d'une enquête est assez complexe et comporte des plans fondamentalement différents pour divers domaines. L'effet de plan des estimations fondées sur l'échantillon total est une somme pondérée des effets de plan selon le domaine. Nous calculons les pondérations sous un modèle approprié et illustrons leur utilisation au moyen de données provenant de l'Enquête sociale européenne (European Social Survey ou ESS).

Mots clés : Stratification; mise en grappes; modèle des composantes de la variance; coefficient de corrélation intraclass; probabilités de sélection.

1. Introduction

En recherche par sondage, l'application de plans de caractéristiques, telles la stratification, la mise en grappes et (ou) l'utilisation de probabilités d'inclusion inégales, qui donnent lieu à des « effets de plan ». L'effet de plan est une mesure qui représente l'effet du plan de sondage sur la variance d'une estimation. Fondé sur le plan de sondage, il est défini comme suit (voir Lohr 1999, page 239) :

$$deff(plan, statistique) = \frac{V(\text{estimation d'après un cas avec le même nombre d'unités d'observation})}{V(\text{estimation d'après le plan d'échantillonnage})}$$

ou cas indique un échantillon aléatoire simple. Le recours à la mise en grappes et (ou) à des probabilités d'inclusion inégales produit habituellement des effets de plan dont la valeur est supérieure à 1,0; autrement dit, la variance d'une estimation est plus grande que celle de l'estimation établie d'après un échantillon aléatoire simple contenant le même nombre d'observations. La prise en compte des effets de plan est très importante lorsqu'on décide d'avancer de la taille de l'échantillon d'une enquête. Ainsi, si l'on prévoit mener une enquête comparative entre plusieurs pays, il est très utile de disposer d'information sur les effets de plan pour ces pays. Il est alors possible de choisir les tailles d'échantillon nettes de façon que la précision des estimations soit approximativement uniforme. Pour cela, la taille d'échantillon qui serait nécessaire sous cas (taille effective d'échantillon) pour obtenir un degré donné de précision doit être multipliée par l'effet de plan prévu.

L'Enquête sociale européenne (ESS, voir www.european-socialsurvey.com) est un programme d'enquête dans lequel les effets de plan sont pris en compte pour le calcul des

tailles nettes d'échantillon, en cherchant à obtenir la même taille effective d'échantillon pour chaque pays ($n_{eff} = 1\,500$). Des 22 pays qui ont participé au premier cycle de l'ESS, trois seulement, le Danemark, la Finlande et la Suède, ont utilisé un plan de sondage avec probabilités de sélection égales, sans mise en grappes (cas). Pour tous les autres, il a fallu prédire l'effet de plan avant l'étude. On peut utiliser, pour cela, une approche fondée sur un modèle (voir Gabler, Häder et Lahiri 1999) qui fait la distinction entre l'effet de plan dû à un échantillonnage avec probabilités d'inclusion inégales (1^{er} terme) et l'effet de plan dû à la mise en grappes (2^e terme) :

$$deff = m \frac{\sum_{i=1}^I m_i w_i^2}{\sum_{i=1}^I m_i w_i} \times \frac{1}{2} \times [1 + (b^* - 1)p] = deff_p \times deff_c \quad (1)$$

où m_i représente les répondants dans la i^{e} classe de probabilités de sélection, chacun recevant un poids de w_i , p est le coefficient de corrélation intragrappe et

$$b^* = \frac{\sum_{c=1}^C \left(\sum_{j=1}^J w_{cj}^2 \right)}{\sum_{c=1}^C \sum_{j=1}^J w_{cj}^2}$$

où b^* est le nombre d'observations dans la grappe c ($c = 1, \dots, C$) et w_{cj} est le poids de sondage de l'élément j dans la grappe c . Il s'agit évidemment d'une simplification reposant sur l'hypothèse qu'il n'existe aucune association entre y et w_{ij} , ou entre w_{ij} et b^* , et ne tenant compte d'aucun effet de stratification, qui aura tendance à être avantageuse et modeste. Voir Lynn, Gabler, Häder et Laaksonen (2007, à paraître), ainsi que Park et Lee (2004) pour une discussion de la sensibilité des prédictions de *deff* à

Gonzalez, M.F., et Hoza, C. (1978). Small area estimation with application to unemployment and housing estimation. *Journal of the American Statistical Association*, 73, 7-15.

Gonzalez, M.F., et Waksberg, J. (1973). Estimation of the errors of synthetic estimates. Article présenté à la première réunion de l'International Association of Survey Statisticians, Vienne, Autriche, 18-25 août.

Levy, P.S. (1971). The use of mortality data in evaluating synthetic estimates. *Dans Proceedings of the American Statistical Association, Social Statistics Section*, 328-331.

Marker, D.A. (1995). *Small area estimation: A Bayesian perspective*. Thèse non publiée, University of Michigan, Ann Arbor, Michigan.

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

Pfaffermann, D. (2002). Small area estimation-New developments and directions. *Revue Internationale de Statistique*, 70, 125-143.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of the mean square error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Purcell, N.J., et Kish, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.

Purcell, N.J., et Kish, L. (1980). Postcensal estimates for local areas (or domains). *Revue Internationale de Statistique*, 48, 3-18.

Rao, J.N.K. (2003a). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rao, J.N.K. (2003b). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2, 145-169.

Schaible, W.L. (1978). Choosing weight for composite estimators for small area statistics. *American Statistical Association*, 741-746.

Schaible, W.L. (1995). Ed. *Lecture Notes in Statistics: Indirect Estimators in U.S. Federal Programs*, New York: Springer.

Singh, M.P., Gambino, J. et Mantel, H.T. (1994). Les petites régions : problèmes et solutions. *Techniques d'enquête*, 20, 3-23.

Thompssen, L., et Holmoy A.M.K. (1998). Combining data from surveys and administrative record system: The Norwegian experience. *Revue Internationale de Statistique*, 66, 201-221.

Chand, N., et Alexander, C.H. (1995). Indirect estimation of rates and rates for small continuous measurement. *Proceeding of the Section on Survey Research Methods*, American Statistical Association, 549-554.

Copas, J.B. (1972). Empirical Bayes methods and the repeated use of a standard. *Biometrika*, 59, 349-360.

Cressie, N. (1989). Empirical Bayes estimation of undercount in the decennial census. *Journal of the American Statistical Association*, 84, 1033-1044.

Ghosh, M., Natarajan, K., Stroud, T.W.F. et Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal (avec discussion). *Statistical Science*, 9, 65-93.

Ghosh, M., et Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London.

Bibliographie

Les travaux de recherche décrits dans l'article ont été financés en partie par le Centre de recherche statistique de l'Iran. Les auteurs remercient les examinateurs et le rédacteur adjoint de leurs nombreux commentaires utiles et constructifs. Mes remerciements sincères vont à Jim Lepkowski pour son aide amicale. Les opinions exprimées sont celles des auteurs et ne reflètent pas forcément celles du Centre statistique d'Iran

Remerciements

générique pour l'Iran. En outre, les méthodes d'estimation pour petits domaines devraient être appliquées non seulement à l'estimation des taux de chômage, mais aussi à celle d'autres paramètres, et les estimations qu'elles produisent devraient être comparées à celles obtenues au moyen de plans d'échantillonnage distincts.

Comme le font remarquer Singh, Gambino et Mantel (1994, page 3),
« On devrait prendre conscience de la question des plans de sondage pour les grandes enquêtes. Les plans d'échantillonnage devraient être conçus de manière que l'on puisse obtenir des données régionales fiables à l'aide d'estimateurs de plan ou de modèle. »
Par conséquent, le CSI doit remanier les plans de sondage afin qu'ils reflètent les besoins des petites régions.

2. Les estimateurs utilisés pour l'estimation pour petits domaines donnent habituellement de bons résultats à mesure qu'augmente la taille d'échantillon. Pour améliorer les estimations provinciales, la taille de l'échantillon national peut être accrue de façon à obtenir des échantillons de plus grande taille pour chaque province. En outre, les provinces ayant des caractéristiques semblables, comme le taux de chômage, les variables sociodémographiques, et ainsi de suite, peuvent être regroupées. Des échantillons de taille distincte seraient alors déterminés pour chaque groupe.

3. L'ajout de variables supplémentaires appropriées, qui sont corrélées à la variable d'intérêt, joue un rôle essentiel dans l'amélioration des estimateurs.
- Une seule variable (âge) a été utilisée dans l'estimateur synthétique pour diviser l'échantillon, mais on peut choisir une autre variable ou une combinaison de variables à cet effet. Les post-strates de l'estimateur synthétique devraient être formées en fonction de variables qui réduisent la variation dans chaque post-strate. Ces variables peuvent influencer indirectement sur l'estimateur composite également. Le modèle EB peut être amélioré en y ajoutant de meilleurs renseignements supplémentaires. Par conséquent, il est important de faire l'essai de diverses variables supplémentaires afin de découvrir le meilleur modèle. Dans les présents travaux, nous n'avons utilisé que la population économique-ment active (PEA) comme variable indépendante dans le modèle, mais d'autres variables pourraient produire de meilleures estimations.

4. L'estimateur composite produit de relativement meilleurs résultats que les estimateurs synthétique et EB. Cependant, notre étude vise uniquement à donner une première idée de l'utilité des méthodes d'estimation pour petits domaines. D'autres travaux seront nécessaires en vue de mettre au point une méthodologie d'estimation pour petits domaines.

Les estimateurs composite et EB donnent habituellement de bons résultats quand le ratio S_g/S_c est égal ou supérieur à 10 % pour une province, parce que les composantes directes des estimateurs (2) et (8) sont relativement stables et reçoivent un poids plus important, particulièrement dans le cas de l'estimateur EB. Les provinces de Téhéran, de Khorasane, de Khuzestan et d'Esfahan sont de ce type, tandis que celles de Bushehr, d'Ilam, de Kohkiluyeh et Buyer Ahmad et de Semnan ne le sont pas.

4. Conclusion

Dans les pays en voie de développement tels que l'Iran, il est fréquent que des dossiers administratifs ne soient disponibles ni pour les petites ni pour les grandes régions. Les enquêtes peuvent produire des estimations satisfaisantes pour les grandes régions, mais non pour les petites. Les recensements périodiques ne permettent pas de fournir toutes les données nécessaires à l'établissement de politiques efficaces et à une bonne planification. Ces limites donnent lieu à des lacunes dans les statistiques officielles. Par conséquent, les activités de planification statistique du Centre statistique d'Iran (SCI) visent à combler ces lacunes en utilisant de nouvelles méthodes et stratégies. Le présent article propose une stratégie rentable pour surmonter certaines de ces limites.
Les résultats de l'étude confirment l'idée qu'un plan d'échantillonnage de portée nationale peut remplacer des plans d'échantillonnage provinciaux distincts si l'on applique les méthodes d'estimation pour petits domaines appropriées. L'échantillon national considère comprend près de 13 000 personnes, tandis que les 21 échantillons provinciaux distincts englobent, en tout, près de 100 000 personnes. Le tirage d'échantillons provinciaux est la méthode utilisée à l'heure actuelle par le CSI pour produire des estimations provinciales. L'utilisation d'un plan de sondage de portée nationale réduirait les coûts de plus de 80 %. En outre, il convient de souligner ce qui suit :

1. Bien que certaines méthodes d'estimation pour petits domaines ne s'appuient pas sur des données d'échantillon existantes provenant de tous les petits domaines (ou petites régions), la stratégie destinée à produire des estimations provinciales sera plus appropriée si les petits domaines d'intérêt sont définis a priori. L'échantillon national peut alors être réparti entre ces petits domaines afin de produire des estimations directes fondées sur le plan de sondage. Il est important d'ajuster le plan de sondage de façon à tenir compte des méthodes d'estimation pour petits domaines avant que la collecte de données ne débute.

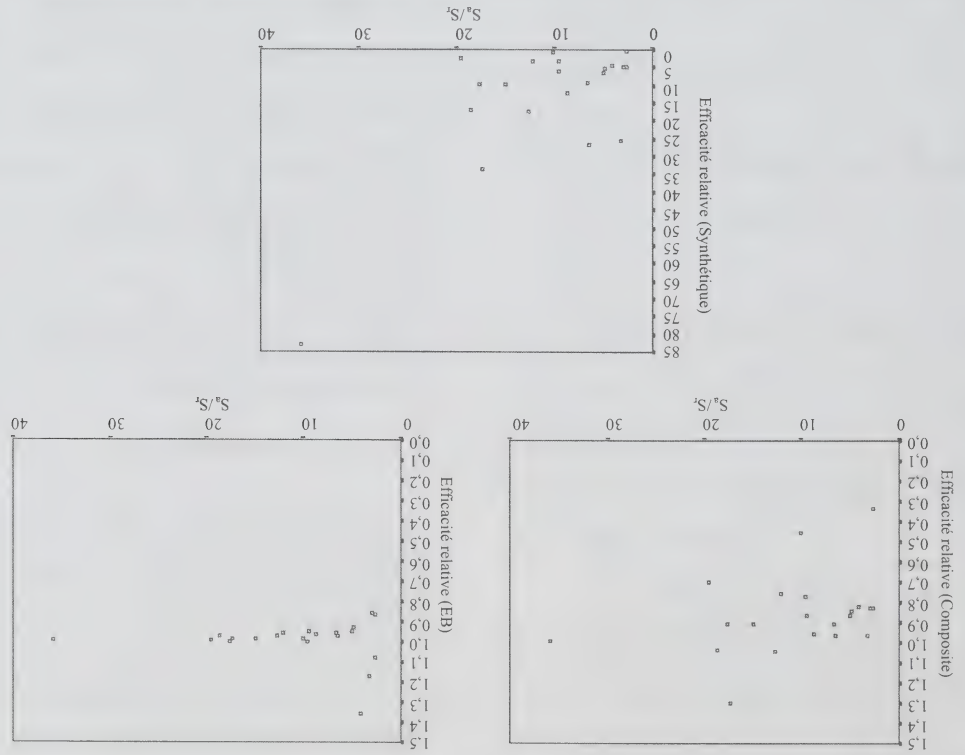


Figure 3. Efficacité relative (EOM estimée de l'estimateur indirect/variance estimée de l'estimateur direct) en fonction du ratio S_a/S_r (une échelle différente a été utilisée sur l'axe vertical pour le diagramme de l'estimateur synthétique pour le rendre plus lisible).

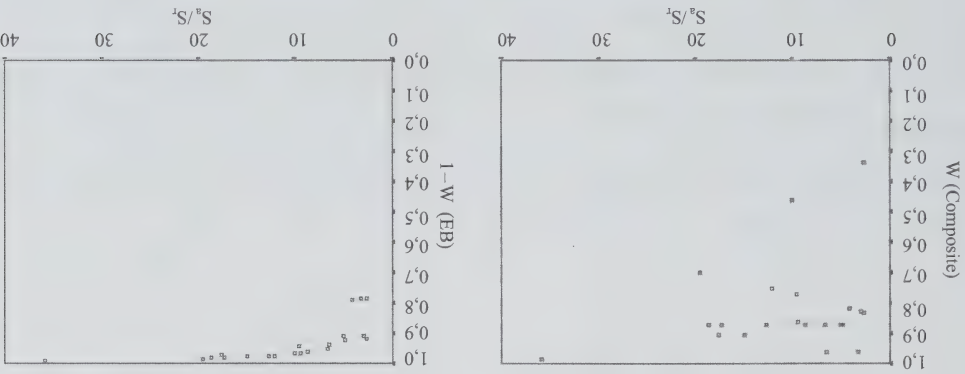
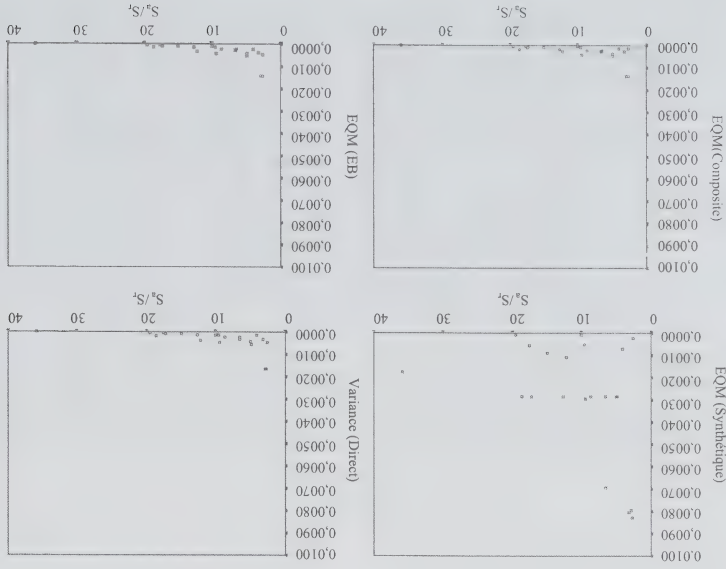


Figure 4. Poids des composantes directes des estimations composite et EB en fonction du ratio S_a/S_r .

Figure 2. EQM des estimations en fonction du ratio S_a/S_p .



plus élevé, respectivement. En général, l'estimateur synthétique produit des résultats médiocres, si l'on choisit pour critères l'EAM, l'EM, l'EQM et l'ER, même si les estimations synthétiques obtenues pour certaines provinces sont, individuellement, plus proches des valeurs réelles que les autres estimations.

Semman et d'Azerbaijan de l'Ouest. Pour l'estimateur composite, Rao (2003a, page 58) déclare que « le poids optimal $M'_{\text{opt}}^{\text{poids}}$ s'approche de zéro ou de un lorsque l'EQM de l'un des estimateurs qui le compose est beaucoup plus grande que celle de l'autre, c'est-à-dire quand la valeur de $f_i = EQM(P_i^1) / EQM(P_i^2)$ est grande ou faible. Dans ces conditions, l'estimateur dont l'EQM est la plus grande ajoute peu d'information et il est donc préférable d'utiliser la composante ayant l'EQM la plus petite. » Ce commentaire est illustré clairement pour les provinces de Bushehr ($W = 0,962355$, $ER^S = 25,27$), de Sistan et Balouchestan ($W = 0,963670$, $ER^S = 26,53$) et de Téhéran ($W = 0,988083$, $ER^S = 83,08$), parce que les estimations directes pour ces provinces ont une EQM plus faible que les estimations ascendantes entre le poids et le ratio S_a/S_i pour l'estimateur EB. Pour l'estimateur composite, le poids le plus faible et le poids le plus élevé correspondent aux provinces ayant le ratio S_a/S_i le plus faible et le plus élevé, respectivement.

Toutefois, les estimations synthétiques ont été calculées dans les conditions les plus défavorables. Les valeurs des ECA appliquées pour les construire sont fondées sur le recensement de 1986 (amietéur de dix ans à l'année pour laquelle les estimations sont produites). En outre, les estimations directes obtenues pour les deux premières post-strates sont assez différentes de celles obtenues pour les autres, ce qui produit de mauvaises estimations synthétiques.

Pour résoudre le premier problème, il faudrait établir des dossiers administratifs, pour le deuxième, l'estimation au niveau des post-strates devrait être prévue lors de l'établissement du plan de sondage. Si l'on tient compte non seulement de l'estimation par post-strate, mais aussi de la classification des provinces lors de l'établissement du plan de sondage, on pourra s'attendre à obtenir de bonnes estimations directes pour les post-strates. Par conséquent, on pourra aussi s'attendre à obtenir de bonnes estimations synthétiques pour les provinces. La classification des provinces peut accroître l'homogénéité grâce au regroupement des provinces semblables par classe et à l'utilisation des données d'échantillon provenant des provinces d'une classe donnée uniquement pour produire les estimations directes par post-strate en vue de construire les estimations synthétiques pour ces provinces.

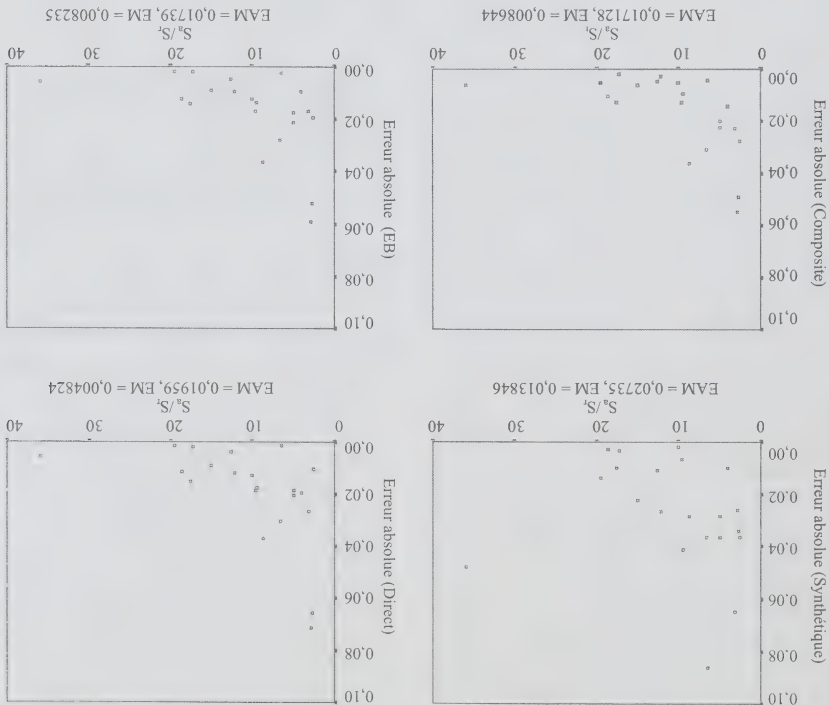


Figure 1. Erreurs absolues des estimation en fonctions du ratio S_a/S_s .

Le ratio S_a/S_s est d'autant plus forte que la dépendance à l'égard des estimations directes provinciales est grande.

Pour toutes les provinces, l'efficacité relative (ER) des trois estimateurs indirects comparativement à l'estimateur direct est souvent inférieure ou égale à l'unité pour les estimateurs composite et EB, et supérieure à l'unité pour l'estimateur synthétique. L'efficacité relative des estimations composites est bonne pour certaines provinces : Seman (0,34), Azerbaijan de l'Ouest (0,46), Khorasan (0,70), Kernanshah (0,75) et Hamadan (0,77). Les moyennes des ER ($ER_s = 13,6$, $ER_c = 0,8595$ et $ER_{EB} = 0,9951$) indiquent que l'estimateur composite est le plus efficace des trois estimateurs indirects. En outre, à la figure 3, à mesure que S_a/S_s augmente, ER_{EB} tend vers un. Comme la figure 2, la figure 3 pourrait être trompeuse pour l'estimateur synthétique.

Il est systématiquement accordé plus de poids à la composante directe, g_i , de l'estimateur donné par l'expression (8) qu'à la composante indirecte. Il en est ainsi pour l'estimateur composite, excepté pour les provinces de

La valeur la plus faible de l'EQM est toujours donnée par l'estimateur composite ou l'estimateur EB (voir la figure 2). Cependant, l'EQM de l'estimateur composite est souvent plus faible que celle de l'estimateur EB. L'EQM de l'estimateur synthétique est systématiquement plus élevée que celle des autres estimateurs, même l'estimateur direct.

À mesure que le ratio S_a/S_s augmente, l'EQM diminue pour les estimateurs direct, composite et EB (voir la tendance décroissante à la figure 2). Cet effet est très important pour Téhéran ($S_a/S_s = 36\%$). De nouveau, les provinces d'Illam et de Kohkiluyeh-o-Boyer Ahmad, qui toutes deux ont une faible population et un ratio S_a/S_s très petit, font exception pour les trois estimateurs. Le nuage de points de la figure 2 pour l'estimateur synthétique pourrait être trompeur, parce que nous avons utilisé l'EQM au lieu de l'EQM pour sept provinces. Cependant, les valeurs pour les quatre provinces mentionnées précédemment (Sistan et Balouchestan, Bushehr, Téhéran et Lorestan) ne concordent pas non plus avec ce nuage de points. En règle générale, pour tout estimateur considéré ici, la relation entre l'EQM et

3.2 Résultats

Nous présenterons les résultats en quatre parties. En premier lieu, nous examinons le biais sous forme de l'erreur et de l'erreur absolue en prenant pour critères l'erreur moyenne (EM) et l'erreur absolue moyenne (EAM). En deuxième lieu, nous comparons les erreurs quadratiques moyennes (EQM) calculées pour les diverses méthodes. En troisième lieu, nous évaluons l'efficacité des estimateurs indirects comparativement à l'estimateur direct. Enfin, nous analysons les poids des composantes directes dans les estimations (2) et (8). Tous les résultats sont illustrés au moyen de figures appropriées, mais des renseignements détaillés sont donnés au tableau 2.

Soit S_a la taille d'échantillon attribuée à une province particulière d'après l'échantillon national et S_p la taille d'échantillon requise individuellement pour la province. Autrement dit, s'il existe un échantillon de taille S_p pour la province, il est possible de calculer une estimation directe pour cette dernière. Par conséquent, $(S_a/S_p) \times 100$ indique dans quelle mesure la taille d'échantillon disponible (S_a) est appropriée pour une province donnée. Cette mesure est utilisée sur l'axe horizontal de tous les diagrammes pour permettre la comparaison des effets de taille d'échantillon.

L'estimateur synthétique est celui dont l'EAM est la plus grande, celle-ci excédant même celle de l'estimateur direct (voir figure 1). Inversement, les EAM des estimateurs composite et EB sont les plus faibles et fort semblables. Si nous choisissons l'erreur moyenne (EM) comme critère, nous constatons que tous les estimateurs surestiment et de la forte population de jeunes dans ces provinces.

Tableau 2
Caractéristiques des provinces et des estimateurs

Province	PEA	S _a	S _p	S _a /S _p	ER ^a	ER ^b	ER ^c	E _A ^c	E _A ^{EB}	E _A ^V	EQM ^c	EQM ^E	EQM ^{EB}	EQM ^V
Bahleyst	133 449	146 450	3,25%	0,96	1,17	25,57	0,0330	0,01687	0,06501	0,02641	0,0003830	0,000417	0,0003813	0,000417
Chahmahal et Bakhtiyan*	141 124	203 463	5,08%	0,87	0,95	6,52	0,02136	0,02135	0,03644	0,02031	0,0003813	0,000417	0,0003813	0,000417
Esfahan	883 653	1032 585	17,69%	0,90	1,00	9,56	0,01268	0,01421	0,00990	0,01504	0,0000533	0,000059	0,0000533	0,000059
Fars	795 175	925 617	12,50%	0,91	0,99	9,69	0,00886	0,00886	0,02235	0,00994	0,0000786	0,000162	0,0000786	0,000162
Gilan	734 966	825 356	12,50%	0,91	0,97	17,25	0,00848	0,00848	0,00460	0,01107	0,0000393	0,0000393	0,0000393	0,0000393
Hamedan	439 517	493 450	9,65%	0,77	1,00	3,36	0,01294	0,01091	0,00675	0,01880	0,0001755	0,0000540	0,0001755	0,0000540
Hormozgan*	168 288	198 406	4,06%	0,84	0,93	5,12	0,01984	0,01701	0,00675	0,01862	0,0004731	0,0000519	0,0004731	0,0000519
Ilam	84 210	111 406	2,77%	0,83	0,87	4,94	0,04901	0,05201	0,03395	0,06579	0,0013919	0,000231	0,0013919	0,000231
Kermanshah	312 768	450 520	8,77%	0,96	0,97	12,00	0,03615	0,03672	0,02864	0,03724	0,0002283	0,000297	0,0002283	0,000297
Kermanshah	357 096	436 575	12,29%	0,75	0,96	3,07	0,00265	0,00288	0,00264	0,01210	0,0002747	0,000349	0,0002747	0,000349
Khorasan	1 410 689	1 587 812	18,59%	0,70	0,99	2,36	0,00193	0,00193	0,01333	0,00169	0,0000298	0,000042	0,0000298	0,000042
Khuzestan*	609 044	786 422	18,69%	1,03	0,97	16,83	0,01034	0,01247	0,00308	0,01140	0,0001760	0,000166	0,0001760	0,000166
Kohgiluyeh-o-Boyer Ahmad	90 655	105 375	2,99%	0,83	0,86	4,83	0,05486	0,05932	0,02630	0,07165	0,0013629	0,000297	0,0013629	0,000297
Kurdistan	276 575	341 520	6,56%	0,91	0,95	9,22	0,03105	0,02814	0,03641	0,03027	0,0002833	0,000297	0,0002833	0,000297
Lorestan	31 918	341 520	9,59%	0,86	0,95	6,22	0,00949	0,01383	0,04101	0,01754	0,0004090	0,000451	0,0004090	0,000451
Mazandaran*	917 259	1 043 601	17,31%	1,30	0,98	33,57	0,00199	0,00188	0,00310	0,00183	0,0001112	0,0000854	0,0001112	0,0000854
Semnan	110 166	121 471	2,69%	0,34	1,08	0,51	0,02776	0,01929	0,03661	0,01042	0,0001534	0,000491	0,0001534	0,000491
Sistan-o-Balouchestan	277 752	2 913 812	35,99%	0,96	0,97	26,53	0,00431	0,00228	0,04767	0,00055	0,0002519	0,000254	0,0002519	0,000254
Tehran	2 343 290	2 913 812	35,99%	0,99	1,00	83,08	0,00605	0,00573	0,00767	0,00028	0,0000209	0,000118	0,0000209	0,000118
Azerbaïdjan-e-Gharbi (de l'ouest)	522 976	654 500	10,14%	0,46	0,98	0,85	0,00505	0,01247	0,00182	0,01309	0,0000529	0,0001024	0,0000529	0,0001024
Yazd	162 892	207 508	4,11%	0,82	1,36	4,52	0,01414	0,00986	0,01008	0,01950	0,0001299	0,000215	0,0001299	0,000215

* Données les provinces pour lesquelles l'expression (3) produit des estimations négatives de l'EAM.
PEA : Population économique active
S_a : Taille d'échantillon requise
ER : Efficacité relative
EA : Erreur absolue
EQM : Erreur quadratique moyenne (la valeur est indiquée en caractères gras pour chaque province)
C, EB, S et D représentent les estimateurs composites, synthétique et direct, respectivement.
Statistique Canada, N° 12-001-XPB au catalogue

3.1 Méthodes de calcul

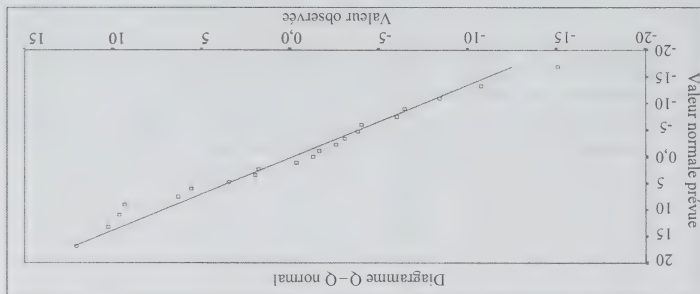
Pour produire des estimations synthétiques, nous avons défini des post-strates en fonction de six groupes d'âge. Au tableau (1), nous présentons, pour chaque groupe, le taux de chômage basé sur l'échantillon national et sa valeur réelle correspondante basée sur le Recensement de 1996, ainsi que les erreurs absolues des estimations.

Tableau 1
Caractéristique des post-strates

Groupe d'âge	Taux estimé (%)	Valeur réelle	Erreur absolue
10-15	0,3240	0,2826	0,0414
16-20	0,2402	0,2629	0,0227
21-25	0,1868	0,1856	0,0012
26-30	0,0811	0,0802	0,0009
31-50	0,0363	0,0366	0,0003
Plus de 50 ans	0,0653	0,0648	0,0005

Les estimations obtenues pour les deux premiers groupes sont entachées d'une très grande erreur. Par conséquent, dans l'expression (1), si une province fournit une grande PEA pour ces groupes d'âge, son estimateur synthétique pourrait ne pas donner de bons résultats. Nous avons utilisé les données du Recensement de 1986 pour calculer les PEA pour toutes les provinces et cellules (N_i et N_{ij} dans l'expression (1)), parce qu'en l'absence de données admnistratives, le recensement le plus rapproché de 1996 est la source principale de données à tout niveau.

Pour construire les estimations composites, nous avons réparti les provinces en deux groupes. Le premier comprend 14 provinces auxquelles nous avons appliqué le poids donné par l'expression (3) et le second, sept provinces auxquelles nous avons appliqué le poids commun $W^C = 0,873184$ fondé sur (6). Comme l'estimateur donné par l'expression



Pour tester la normalité, nous avons utilisé un diagramme quantile-quantile (Q-Q plot) normal et un test de Shapiro-Wilk pour les résidus standardisés du modèle ajusté. Les points du diagramme Q-Q sont proches d'une droite et le test ne mène pas au rejet de l'hypothèse nulle de normalité (valeur $p = 0,851$).

$$t^2 = 0.5596389, \hat{\beta} = \begin{pmatrix} -2.066874 \\ -1.273 \times 10^{-7} \end{pmatrix}$$

Les valeurs estimées de t^2 et $\hat{\beta}$ sont

$$X = \begin{pmatrix} 1 & 133 & 449 \\ 1 & 141 & 124 \\ 1 & 883 & 653 \\ 1 & 795 & 714 \\ : & : & : \\ 1 & 522 & 976 \\ 1 & 162 & 892 \end{pmatrix}$$

les PEA, respectivement :

première et deuxième colonnes contiennent des valeurs l et expérience de dimensions 21×2 suivante, dont les équations (11), nous avons utilisé la matrice de plan moment et avons obtenu $t^2 = 0,3117194$. Pour résoudre les auteurs). Ce programme nécessite une estimation initiale de t^2 que nous avons calculée par la méthode d'estimation du SAS/IML (ce programme peut être obtenu auprès des de Prasad et Rao (1990) en utilisant un programme d_i^2 en utilisant la méthode delta, puis t^2 suivant la méthode

erreurs quadratiques moyennes (EQMM). Pour construire les estimations EB, nous avons estimé ces sept provinces, nous avons utilisé la moyenne des (4) produit des estimations négatives de $EQM(\hat{p}_i^s)$ pour

L'EQM est positive (voir Gonzalez et Waksberg (1973) pour plus de précisions) :

$$W^C = \frac{\left(\sum_{i=1}^I V(p_d^i) / \sum_{i=1}^I EQM(p_s^i) \right) + 1}{1} \quad (6)$$

ce poids repose sur les estimations de t^2 et de d_t^2 . Par application de la méthode delta, $(g_t^i)'V(p_d^i)$ est la dérivée première de $g_t^i = Ln(p_d^i/1 - p_d^i)$. En nous inspirant de Chand et Alexander (1995), nous obtenons les estimations de β et de t^2 en résolvant simultanément

$$\begin{cases} t^2 = (\mathbf{g}' - \mathbf{X}\beta)'V^{-1}(\mathbf{g} - \mathbf{X}\beta) / (I - k) \\ \beta = (\mathbf{X}'V^{-1}\mathbf{X})^{-1}\mathbf{X}'V^{-1}\mathbf{g} \end{cases} \quad (11)$$

où $V = \text{Diag}(d_1^2 + t^2, \dots, d_I^2 + t^2)$. Soulignons que les équations (11) sont résolues par itération numérique en partant d'une valeur initiale pour t^2 .

Les estimateurs EB et composite présentent des similarités, bien qu'ils soient obtenus en suivant des approches différentes. L'un et l'autre ont deux composantes, c'est-à-dire une composante directe (p_d^i dans (2) et g_t^i dans (8)) calculée d'après des données d'échantillon provincial et une composante indirecte (p_s^i dans (2) et $x_t^i\beta$ dans (8)) construite d'après les données de l'échantillon national et (8) accordent tous deux plus de poids à la composante indirecte lorsqu'elle est fiable. Sinon, la composante directe reçoit plus de poids. Des précisions supplémentaires sont données dans Cressie (1989), Ghosh et coll. (1998) et Rao (2003 a, b).

3. Estimation pour l'Iran

Nous avons produit des estimations pour 1996, parce que les taux réels de chômage de 1996 sont connus pour chaque province grâce au Recensement de 1996. Par conséquent, il est possible de calculer le biais réel pour chaque estimation.

En 1996, le pays était constitué de 26 provinces. Cependant, nous n'étudions que 21 ici, parce que les renseignements supplémentaires provenant du Recensement de 1986 n'étaient disponibles que pour 21 provinces dont les limites géographiques n'ont pas changé entre 1986 et 1996. national, nous avons établi un plan de sondage en déterminant la taille d'échantillon nécessaire pour estimer le taux de chômage pour le pays dans son ensemble. Chaque province représente un petit domaine. L'échantillon national a été réparti entre les 26 provinces proportionnellement à la population de ces dernières afin de disposer de données d'échantillon provenant de chaque province (approche descendante). Cela nous a permis de calculer des estimations directes fondées sur le plan de sondage pour chaque province et les variances correspondantes requises pour les estimateurs EB et composite. Le plan de sondage permet de produire de bonnes estimations pour le pays dans son ensemble et pour certaines provinces.

où

$$\mathbf{g}' = (Ln \frac{p_d^1}{1 - p_d^1}, \dots, Ln \frac{p_d^I}{1 - p_d^I})$$

2.3 Estimateur empirique bayésien (EB)

Plus d'attention a été accordée aux méthodes d'estimation pour petits domaines fondées sur un modèle qu'aux estimateurs synthétique et composite. Marker (1999) considère que les méthodes d'estimation pour petits domaines ont un élément commun exprimé au moyen de modèles de régression. La méthode EB rentre dans la catégorie des méthodes de régression. Considérons le modèle mixte suivant (voir Rao (2003a, page 76)) :

$$\mathbf{g} = \mathbf{X}\beta + \mathbf{v} + \mathbf{e} \quad (7)$$

X est une matrice de plan d'expérience de dimensions $I \times k$ de variables supplémentaires, β est un vecteur de dimensions $k \times 1$ de paramètres inconnus, et \mathbf{v} et \mathbf{e} sont des vecteurs aléatoires de dimensions $I \times 1$ (I est le nombre de provinces). Supposons que :

1. \mathbf{v} et \mathbf{e} sont indépendants;
2. $E(\mathbf{e}) = 0$ et $\text{Var}(\mathbf{e}) = \text{Diag}(d_1^2, \dots, d_I^2)$;
3. $\mathbf{v} \sim N(0, \Sigma)$, où $\Sigma = \text{Diag}(t^2, \dots, t^2)$.

Ghosh et Meeden (1997) montrent que l'estimation EB du t^{e} élément de \mathbf{g} est :

$$\hat{g}_{EB}^i = \hat{W}_i' x_t^i \beta + (1 - \hat{W}_i') g_t^i \quad (8)$$

où x_t^i et g_t^i sont la t^{e} ligne et la t^{e} composante de X et \bar{g} respectivement, et \hat{W}_i' est une estimation de

$$W_i' = \frac{p_d^i}{d_t^2 + t^2} \quad (9)$$

Par conséquent, l'estimation EB du t^{e} taux est :

$$\hat{p}_{EB}^i = \frac{\exp(\hat{W}_i' x_t^i \beta + (1 - \hat{W}_i') g_t^i)}{1 + \exp(d_t^i p_d^i / (1 - p_d^i) g_t^i)} \quad (10)$$

Il est évident que l'utilisation de (10) requiert deux estimations, celles de β et du poids donné par (9). Par ailleurs,

où P_i est une estimation directe fondée sur le plan de sondage du taux de chômage dans la post-strate j_i , N_i est la PEA de la province i et N_{ij} est la PEA dans l'intersection de la province i et de la post-strate j_i , c'est-à-dire la cellule (i, j_i) . L'estimation synthétique pour la i^{e} province est calculée conformément à la définition officielle du taux de chômage en Iran.

L'estimation synthétique s'appuie sur toutes les données de l'échantillon national grâce à l'utilisation des estimations directes nationales du taux de chômage d'après les post-strates. Elle est basée sur les six estimations du taux de chômage par « post-strate » calculées sur l'ensemble des provinces, plutôt que sur les estimations spécifiques des six « cellules ». Par conséquent, ce processus revient à **en-punir de la force (information)**, puisque chaque province contribue à l'échantillon national grâce au regroupement des unités d'échantillonnage provinciales en vue de surmonter les problèmes posés pour chaque province par la petite taille de l'échantillon.

Cet estimateur a trois limites :

1. L'estimateur synthétique donne des résultats d'autant meilleurs que la variation entre les post-strates est faible. Autrement dit, le taux de chômage dans chaque groupe d'âge devrait être à peu près le même dans toutes les provinces. L'utilisation des estimations directes nationales par post-strate de façon uniforme pour toutes les provinces n'est admissible que sous cette hypothèse. Si l'hypothèse d'homogénéité n'est pas satisfaite, l'estimateur synthétique ne peut pas refléter les variations particulières au niveau des petits domaines et les estimations pourraient être gravement biaisées.

2. S'il existe plusieurs variables importantes pour la post-stratification, il est fréquent qu'on ne puisse pas les utiliser toutes dans l'estimateur synthétique, parce que la taille des échantillons des post-strates (après recoupement de plusieurs variables) est trop faible et produit des estimations directes inadéquates au niveau de la post-strate. Généralement, un grand nombre de post-strates donne lieu à des estimations directes de mauvaise qualité pour certaines post-strates. Cette situation peut créer de sérieux problèmes lors de l'estimation synthétique, si elle est associée à une grande PEA dans une cellule.
3. La qualité des estimations de la PEA peut avoir une incidence sur les estimations synthétiques. Étant donné le manque de sources de données à jour, comme les dossiers administratifs, des estimations périodiques de la PEA d'après les données du

2.2 Estimateur composite

Recensement de 1986 sont utilisées ici pour produire les estimations synthétiques pour 1996.

L'estimateur composite pour la i^{e} province est une combinaison des estimateurs synthétique et direct pour cette province, à savoir

$$P_i^c = W_i P_i^d + (1 - W_i) P_i^s \quad (2)$$

où P_i^d est l'estimateur direct fondé sur le plan de sondage pour la i^{e} province et $0 \leq W_i \leq 1$. L'expression (2) est une amélioration de l'expression (1) grâce à l'exploitation des deux estimateurs. Autrement dit, dans l'estimateur composite, les écarts interprovinciaux sont pris en compte au moyen des estimations provinciales directes sans biais et l'instabilité de l'estimateur direct est réduite au moyen de l'estimateur synthétique.

Le poids W_i peut être spécifié de façon à réduire au minimum l'erreur quadratique moyenne de P_i^c , $E\hat{Q}M(P_i^c)$. Si l'on suppose que $\text{Cov}(P_i^d, P_i^s) \approx 0$, l'expression du poids se simplifie comme suit

$$W_i^{\text{opt}} = \frac{V(P_i^d) / E\hat{Q}M(P_i^s)}{V(P_i^d) / E\hat{Q}M(P_i^s) + 1} \quad (3)$$

où $V(P_i^d)$ et $E\hat{Q}M(P_i^s)$ sont la variance de P_i^d et l'erreur quadratique moyenne de P_i^s , respectivement. Dans l'expression (3), les poids des estimateurs direct et synthétique figurent dans (2) sont proportionnels aux $E\hat{Q}M$ des deux estimateurs. Voir Schauble (1978) et Rao (2003a, page 58) pour les propriétés de l'estimateur et du poids.

En pratique, nous devrions estimer $E\hat{Q}M(P_i^s)$ et $V(P_i^d)$ pour générer une estimation du poids (3). S'il existe des données d'échantillon provenant de la i^{e} province, d'après le plan de sondage, nous pouvons calculer un estimateur fondé sur le plan de sondage sans biais de $V(P_i^d)$ en utilisant uniquement les données de l'échantillon. Par conséquent, un seul estimateur est nécessaire pour $E\hat{Q}M(P_i^s)$. Sous l'hypothèse que $\text{Cov}(P_i^d, P_i^s) \approx 0$, Ghosh et Rao (1994) ont proposé l'estimateur sans biais

$$E\hat{Q}M(P_i^s) = (P_i^s - P_i^d)^2 - V(P_i^d) \quad (4)$$

Sous la même hypothèse, il est facile de montrer que

$$E\hat{Q}M(P_i^c) = W_i^2 V(P_i^d) + (1 - W_i)^2 E\hat{Q}M(P_i^s) \quad (5)$$

L'estimateur (4) produit parfois des estimations négatives pour certaines provinces, de sorte que le poids donné par l'expression (3) n'est plus calculable. Dans ce cas, au lieu de l'expression (3) et (4), nous avons utilisé, respectivement, le poids combiné donné par (6) et $E\hat{Q}M = (1/I) \sum_{i=1}^I E\hat{Q}M(P_i^s)$, où I' est le nombre de petits domaines dont l'estimation de

méthodes d'estimation pour petits domaines ont été appliquées pour produire des estimations indirectes pour chaque province. Troisièrement, les estimations indirectes ont été évaluées par comparaison aux valeurs réelles correspondantes, en se basant sur le calcul de l'erreur quadratique moyenne (EQM), de l'erreur absolue moyenne (EAM) et de l'erreur moyenne (EM).

Outre cette introduction, l'article contient trois autres sections. À la section 2, nous passons brièvement en revue les trois estimateurs utilisés, y compris les méthodes d'estimation, les EQM correspondantes et les propriétés des estimateurs. À la section 3, nous présentons les estimations et les méthodes de calcul correspondantes, et nous essayons d'évaluer les propriétés des estimateurs. À la section 4, nous présentons nos conclusions en ce qui concerne les estimateurs et les avantages de la stratégie d'estimation pour petits domaines, et nous formulons des recommandations.

2. Aperçu des estimateurs

Nous nous contions de présenter brièvement les estimateurs indirects utilisés pour l'étude. Cependant, le lecteur trouvera une excellente discussion des méthodes d'estimation pour petits domaines dans Rao (2003a). Nous examinons d'abord l'estimateur synthétique, puis l'estimateur composite. Nous considérons aussi l'estimateur empirique bayésien (EB) à titre d'estimateur fondé sur un modèle.

2.1 Estimateur synthétique

Il existe une famille d'estimateurs sur petits domaines qui sont qualifiés de synthétiques, voir Rao (2003a, chapitre 4). Nous décrivons ici celui qui est le plus classique et le plus simple. Pour cet estimateur,

1. le pays est divisé en six post-strates en fonction de six groupes d'âge (voir tableau 1);
2. puis, le nombre de chômeurs est estimé dans chaque province, ce qui donne le numérateur de l'expression (1);
3. enfin, l'estimateur synthétique pour la i^{e} province est obtenu en divisant le nombre estimé de chômeurs dans la province i par la population économiquement active (PEA) de la province, c'est-à-dire

$$\hat{P}_s^i = \left(\sum_{j=1}^6 N_j^i \hat{P}_j^i \right) / N_i \quad (1)$$

sélectionne pour chaque province afin d'estimer les taux de chômage provinciaux. Le taux de chômage pour l'ensemble du pays était ensuite calculé par combinaison pondérée des estimations provinciales. La demande croissante d'estimations du taux de chômage au niveau provincial sur une base mensuelle, ou du moins saisonnière, et le manque de dossiers administratifs en Iran au niveau tant régional que national ont persuadé le CSI d'essayer d'adopter les méthodes d'estimation pour petits domaines comme élément central d'une stratégie révisée en vue de répondre aux besoins des provinces.

La stratégie révisée consiste à concevoir un plan de sondage uniquement au niveau national et à produire des estimations provinciales par des méthodes d'estimation pour petits domaines. Dans cette stratégie, une province représente un petit domaine. Cette stratégie requiert un échantillon de plus petite taille que celui résultant de l'agrégation des échantillons provinciaux. Si la stratégie révisée s'avère applicable en pratique, il sera possible de réduire la durée et le coût de la collecte des données et de produire des estimations provinciales mensuellement. Le plus petit échantillon est plus facile à gérer sur le terrain et les estimations sont moins affectées par les erreurs non dues à l'échantillonnage.

Le présent article vise à répondre aux questions suivantes :

1. Un échantillon national peut-il remplacer les échantillons provinciaux distincts pour estimer les taux de chômage provinciaux?
2. Parmi les trois méthodes d'estimation pour petits domaines, à savoir les estimateurs synthétique, composite et empirique bayésien, laquelle produit les meilleures estimations?

Afin de répondre empiriquement à ces deux questions, nous avons produit des estimations pour 1996, année pour laquelle les valeurs réelles des taux de chômage provinciaux sont connues grâce au Recensement de 1996. Par conséquent, nous pouvons calculer le biais réel de chaque estimation provinciale.

Le processus comprend les trois étapes qui suivent. Premièrement, un échantillon de 13 000 unités a été sélectionné pour l'ensemble du pays (le fichier de données du Recensement de 1996). La taille de l'échantillon a été déterminée au niveau national, puis répartie entre les provinces proportionnellement à leur population. La répartition fournit pour chaque province un échantillon qui permet d'estimer directement le taux de chômage provincial. Les estimations directes ne sont pas nécessairement acceptables pour toutes les provinces, à cause des erreurs d'échantillonnage importantes dues à la petite taille de l'échantillon pour certaines provinces. Deuxièmement, trois

Rivest, L.P., et Vandal, N. (2002). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, 10-13 juillet, 2002, Ottawa, Canada.

Wang, J., et Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y., et Rao, J.N.K. (2000). Estimation bayésienne hiérarchique des moyennes pour petites régions à l'aide de modèles à plusieurs niveaux. *Techniques d'enquête*, 26, 197-206.

You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 1, 3-15.

You, Y., Rao, J.N.K. et Dick, P. (2004). Benchmarking hiérarchique Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.

You, Y., Rao, J.N.K. et Gambino, J. (2003). Estimation du taux de chômage fondée sur un modèle pour l'Enquête sur la population active du Canada : Une approche bayésienne hiérarchique. *Techniques d'enquête*, 29, 27-36.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rao, J.N.K. (1999). Quelques progrès concernant l'estimation régionale fondée sur un modèle. *Techniques d'enquête*, 25, 199-212.

Gelfand, A.E., et Smith, A.F.M. (1991). Gibbs sampling for marginal calculating marginal densities. *Journal of the American Statistical Association*, 85, 972-985.

Gelfand, A.E., et Smith, A.F.M. (1990). Sample-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 74, 268-277.

Fay, R.E., et Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 94, 1074-1082.

Datta, G.S., Lahiri, P., Maiti, T. et Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

4. Conclusion et travaux futurs

Dans le présent article, nous avons étudié le modèle bien connu de Fay-Herriot dans les situations où il est supposé que σ^2_i , la variance d'erreur d'échantillonnage est inconnue et estimée au moyen de l'estimateur sans biais s^2_i , en utilisant l'approche hiérarchique bayésienne. L'approche HB complétée avec échantillonnage de Gibbs tient compte automatiquement de l'incertitude supplémentaire associée à l'estimation de σ^2_i . Nous avons appliqué l'approche HB à l'analyse de deux ensembles de données d'enquête. Nos résultats montrent que l'approche HB proposée sous le modèle I donne d'assez bons résultats, que les tailles des échantillons régionaux soient de grande ou de petite taille. Lors de futurs travaux, l'approche de modélisation HB proposée pourrait être étendue aux modèles de niveau régional généraux étudiés par You et Rao (2002). Les applications de la nouvelle approche de modélisation HB comprennent l'estimation du sous-dénombrement au recensement décrite dans You, Rao et Dick (2004). Sous le modèle I, il est possible d'obtenir les estimateurs HB des variances d'échantillonnage σ^2_i . Ces estimateurs HB de σ^2_i peuvent alors être utilisés comme estimateurs lissés de σ^2_i dans les modèles d'échantillonnage. Les applications et évaluations des estimateurs HB des variances d'échantillonnage comprennent l'estimation du sous-dénombrement au recensement et l'estimation du taux de chômage dans le cadre de l'Enquête sur la population active (EPA) du Canada (You, Rao et Gambino 2003). Nous prévoyons aussi comparer l'approche HB à l'approche EBLUP telle qu'elle a été étudiée par Rivest et Vandal (2002), ainsi que par Wang et Fuller (2003).

Remerciements

Les auteurs tiennent à remercier deux examinateurs, un rédacteur adjoint, le rédacteur en chef délégué et le rédacteur en chef M.P. Singh, de leurs suggestions et commentaires constructifs. Les auteurs remercient aussi J.N.K. Rao, de l'Université Carleton, pour ses suggestions utiles, ainsi que Jack Gambino et Eric Rancourt, de Statistique Canada, pour leurs commentaires au sujet de la première version de l'article. Ces travaux ont été financés grâce aux ressources de financement global de la recherche de la Direction de la Méthodologie de Statistique Canada.

Bibliographie

Arora, V., et Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

appropriées vagues (résultats non présentés ici). Puisque les données sur le lait proviennent d'échantillons de grande taille, nous pouvons également utiliser des lois a priori uniformes sur les composantes de la variance pour les analystes sous le modèle I. Nous obtenons donc les estimations HB fondées sur les lois a priori uniformes et les comparons aux estimations HB fondées sur les lois a priori GI vagues. Ces estimations HB sont presque identiques et stables, l'écart relatif variant de 0,07 % à 2,23 %, avec une valeur moyenne de 0,69 % sur 43 régions, ce qui indique que les estimations a posteriori des moyennes de petite région fondées sur le modèle I sont très stables et ne sont pas sensibles au choix des lois a priori uniformes ni des lois a priori GI vagues, à condition que les tailles d'échantillon et le nombre de petites régions soient relativement grands.

Tableau 5

Comparaison des estimations des moyennes a posteriori pour les données sur le maïs

Comité	GI(a_i, b_i), $a_i = b_i$		Matis	
	0,0001	0,001	0,001	0,01
Franklin	142,862	142,593	143,155	144,311
Pocahontas	91,560	91,912	91,422	91,974
Winnebago	113,130	113,068	121,578	114,430
Wright	123,547	124,170	125,103	125,351
Webster	97,856	98,231	99,132	98,511
Hancock	123,478	123,858	124,395	124,138
Kossuth	114,910	115,281	115,316	115,528
Hardin	135,178	134,157	135,223	136,001
Soya				
Franklin	88,186	89,368	89,145	89,513
Pocahontas	109,052	109,571	107,745	108,176
Winnebago	88,053	87,478	86,267	87,302
Wright	105,825	106,712	109,835	110,252
Webster	109,455	108,392	109,835	110,252
Hancock	102,876	103,413	102,240	101,808
Kossuth	101,862	101,159	101,379	100,808
Hardin	93,397	94,713	93,576	94,767

Tableau 6
Comparaison des c.v. a posteriori pour les données sur le maïs

Comité	GI(a_i, b_i), $a_i = b_i$		Matis	
	0,0001	0,001	0,001	0,01
Franklin	0,129	0,124	0,128	0,125
Pocahontas	0,556	0,351	0,347	0,341
Winnebago	0,311	0,314	0,321	0,324
Wright	0,250	0,246	0,235	0,236
Webster	0,307	0,292	0,285	0,280
Hancock	0,145	0,148	0,148	0,142
Kossuth	0,109	0,110	0,107	0,104
Hardin	0,176	0,173	0,178	0,168
Soya				
Franklin	0,239	0,233	0,231	0,227
Pocahontas	0,276	0,281	0,271	0,296
Winnebago	0,214	0,193	0,196	0,198
Wright	0,232	0,223	0,231	0,226
Webster	0,236	0,231	0,237	0,228
Hancock	0,169	0,165	0,168	0,161
Kossuth	0,148	0,145	0,142	0,135
Hardin	0,217	0,215	0,213	0,213

sur σ^2_ϵ et σ^2_η , les lois conditionnelles complètes pour σ^2_ϵ et σ^2_η sont données par

$$[\sigma^2_\eta | y, \theta, \beta, \sigma^2_\epsilon] \sim \text{GI} \left(\frac{d_\eta - 1}{2}, \frac{(y_i - \theta_i)^2 + d_\eta s^2_\eta}{2} \right),$$
$$[\sigma^2_\epsilon | y, \theta, \beta, \sigma^2_\eta] \sim \text{GI} \left(\frac{m - 2}{2}, \frac{1}{m} \sum_{i=1}^m (\theta_i - x'_i \beta)^2 \right).$$

En postulant une loi a priori uniforme $\pi(\sigma^2_\epsilon) \propto 1$, nous obtenons

$$\pi(\sigma^2_\epsilon | s^2_\eta) \sim \text{GI} \left(\frac{d_\eta}{2} - 1, \frac{d_\eta s^2_\eta}{2} \right),$$

premier lieu, nous pouvons obtenir la loi a posteriori de σ^2_ϵ , connaissant l'estimation directe s^2_η de cette dernière, sous la forme

$$\pi(\sigma^2_\epsilon | s^2_\eta) \propto f(s^2_\eta | \sigma^2_\eta) \cdot \pi(\sigma^2_\eta) \propto (\sigma^2_\eta)^{-d_\eta/2} \cdot \exp\{-\sigma^2_\eta d_\eta s^2_\eta / 2\} \cdot \pi(\sigma^2_\eta).$$

Tableau 4 Comparaison des estimations HB pour les données sur le lait

Petite région	σ^2_ϵ connue ($\sigma^2_\eta = s^2_\eta$)	θ_{HB}	c.v.	e.-l.	σ^2_ϵ inconnue
1	1.020	0.113	0.111	1.021	0.109
2	1.045	0.072	0.069	1.045	0.068
3	1.065	0.073	0.069	1.065	0.069
4	0.767	0.095	0.124	0.770	0.125
5	0.849	0.096	0.113	0.852	0.113
6	0.975	0.103	0.106	0.975	0.102
7	1.058	0.125	0.118	1.055	0.118
8	1.097	0.099	0.090	1.096	0.099
9	1.219	0.122	0.121	1.215	0.121
10	1.192	0.122	0.102	1.190	0.102
11	0.793	0.094	0.119	0.799	0.122
12	1.213	0.131	0.108	1.209	0.107
13	1.206	0.112	0.093	1.203	0.112
14	0.984	0.107	0.109	0.987	0.109
15	1.187	0.105	0.088	1.187	0.104
16	1.156	0.104	0.090	1.156	0.102
17	1.225	0.101	0.083	1.225	0.100
18	1.284	0.115	0.089	1.281	0.113
19	1.234	0.101	0.082	1.235	0.100
20	1.233	0.110	0.089	1.233	0.110
21	1.092	0.097	0.089	1.095	0.089
22	1.192	0.128	0.107	1.193	0.106
23	1.122	0.103	0.092	1.125	0.103
24	1.221	0.113	0.092	1.220	0.111
25	1.193	0.086	0.072	1.193	0.086
26	0.761	0.091	0.120	0.762	0.120
27	0.763	0.092	0.120	0.762	0.119
28	0.734	0.125	0.170	0.732	0.123
29	0.768	0.085	0.110	0.767	0.110
30	0.615	0.076	0.124	0.618	0.123
31	0.769	0.122	0.158	0.767	0.156
32	0.795	0.119	0.150	0.792	0.148
33	0.771	0.091	0.118	0.770	0.117
34	0.612	0.060	0.099	0.613	0.100
35	0.701	0.085	0.121	0.701	0.120
36	0.757	0.094	0.123	0.759	0.123
37	0.534	0.080	0.150	0.538	0.151
38	0.744	0.096	0.129	0.743	0.128
39	0.754	0.082	0.108	0.753	0.108
40	0.768	0.088	0.115	0.768	0.088
41	0.747	0.071	0.095	0.747	0.070
42	0.801	0.093	0.116	0.800	0.092
43	0.682	0.094	0.139	0.682	0.094

L'application de l'échantillonneur de Gibbs sous les lois a priori uniformes est également simple. Cependant, les lois a priori uniformes sur σ^2_ϵ et σ^2_η peuvent mener à des lois a posteriori, ou postérieurs, inappropriées si les tailles d'échantillon et le nombre de petites régions sont faibles. Pour mieux visualiser le problème des lois a priori sur σ^2_ϵ , nous pouvons étudier le modèle 1 en deux étapes. En

à condition que $d_\eta > 2$, ou $n_i > 3$. Alors, nous pouvons utiliser cette loi a posteriori GI appropriée $\pi(\sigma^2_\epsilon | s^2_\eta)$ en tant que loi a priori informative sur σ^2_ϵ dans le modèle d'échantillonnage $y_i | \theta_i, \sigma^2_\epsilon \sim \text{ind } N(\theta_i, \sigma^2_\epsilon)$. Pour les données modifiées sur le maïs et le soja, l'utilisation des lois a priori uniformes sur σ^2_ϵ produira une loi a posteriori incorrecte, comme cela est fait habituellement en pratique lors de l'estimation HB sur petites régions (par exemple, Arora et Lahiri 1997; Datta, Lahiri, et Lu 1999; You et Rao 2000; Rao 2003). Par conséquent, nous n'avons pas à craindre que certaines lois a posteriori soient inappropriées, puisque l'inférence HB correcte devrait être fondée sur des lois a posteriori appropriées. Sous le modèle 2 avec variance d'échantillonnage connue donnée par $\sigma^2_\epsilon = s^2_\eta$, et l'utilisation d'une loi a priori uniforme $\pi(\sigma^2_\epsilon) \propto 1$ sur σ^2_ϵ , la loi a posteriori de σ^2_ϵ sera appropriée à condition que $m > p + 2$, où m est le nombre de petites régions et p est la taille des paramètres de régression β (Rao 2003, page 238). Puisque le nombre de petites régions est habituellement assez grand, cette condition est en général satisfaite en pratique.

Pour l'analyse de sensibilité des lois a priori appropriées vagues, nous pouvons tester la sensibilité des estimations a posteriori au choix des paramètres a priori a_i , b_i ($0 \leq i \leq m$). Sous le modèle 1, nous fixons $a_i = b_i$ à quatre valeurs différentes, c'est-à-dire 0,0001, 0,001, 0,01 et 0,1. Le tableau 5 donne les estimations des moyennes a posteriori pour les données sur le maïs et le soja, et le tableau 6, les c.v. correspondants.

Il est évident, si l'on examine les tableaux 5 et 6, que les estimations a posteriori et les c.v. correspondants sont à peu près les mêmes et stables, ce qui indique que les estimations HB ne sont pas sensibles au choix des lois a priori appropriées vagues. Dans le cas des données sur le lait, les estimations HB sont très stables au choix de ces lois a priori

de la variabilité supplémentaire due à l'estimation de σ^2_{θ} . En moyenne, l'accroissement des e.-t. et des c.v. est de l'ordre de 20 % (ce calcul exclut le comté de Franklin pour les données sur le maïs). Les résultats confirment que si l'on suppose que $\sigma^2_i = s^2_i$, l'estimation directe connue de σ^2_i on obtient une sous-estimation de l'erreur-type et du coefficient de variation de θ_i . L'examen des comtés de Franklin et de Webster pour les données sur le maïs et du comté de Winnebago pour les données sur le soja établit que, dans certains cas où les erreurs d'échantillonnage sont, par hasard, assez faibles, cette sous-estimation est importante.

Tableau 3
Comparaison des estimations HB pour les données sur les cultures

Comté	σ^2_i connue ($\sigma^2_i = s^2_i$)		σ^2_i inconnue	
	θ_{HB}	e.-t.	θ_{HB}	e.-t.
Franklin	155,788	6,061	0,039	142,862
Pocahontas	100,813	28,297	0,281	91,560
Winnebago	113,337	28,406	0,246	113,130
Wright	131,630	28,345	0,215	123,347
Webster	109,030	20,634	0,189	97,856
Hancock	121,682	15,656	0,129	123,478
Kossuth	117,180	11,180	0,097	109,410
Hardin	135,626	23,228	0,171	135,178
Franklin	75,375	16,272	0,216	88,186
Pocahontas	116,943	27,031	0,231	109,052
Winnebago	87,525	10,304	0,118	88,053
Wright	104,184	23,671	0,227	105,825
Webster	115,510	20,789	0,180	109,455
Hancock	101,368	15,741	0,155	102,876
Kossuth	102,388	14,948	0,146	101,862
Hardin	87,455	17,774	0,203	93,397

La comparaison des estimations HB sous les modèles 1 et 2 aux estimations directes peut se faire en se servant des c.v. présentes aux tableaux 1 et 3. Sous le modèle 2, les estimations HB ont un c.v. plus petit que les estimations directes pour six des huit comtés pour les données sur le maïs et, de même, dans six des huit comtés pour les données sur le soja. Dans le cas des deux cultures, pour les deux comtés restants, les c.v. sous le modèle 2 sont les mêmes ou légèrement plus grands que les c.v. des estimations directes par sondage. Par conséquent, les estimateurs provenant du modèle 2 semblent être plus efficaces que les estimateurs directs par sondage. L'examen des estimations directes sous le modèle 1 et des estimations directes par sondage produit des résultats mixtes pour les ensembles de données sur le maïs et le soja. Le modèle 1 tient compte de l'incertitude supplémentaire due à l'estimation des variances d'échantillonnage et, par conséquent, les estimations HB ne sont meilleures dans le cas des données sur le maïs que pour quatre des huit comtés. Dans le cas des données sur le soja, les estimations HB représentent une amélioration par rapport aux c.v. des estimations directes par sondage pour cinq des huit comtés. Pour les autres, les c.v. des estimations

3.3 Lois a priori et analyse de sensibilité

Dans le modèle 1, nous supposons que les variances d'échantillonnage σ^2_i sont indépendantes et suivent une loi a priori gamma inverse $GI(a_i, b_i)$, et que la variance sous le modèle σ^2_i suit aussi une loi a priori gamma inverse $GI(a_0, b_0)$, où a_i, b_i ($0 < i \leq m$) sont des constantes connues fixées à une valeur très faible afin de refléter les connaissances vagues au sujet de σ^2_i et σ^2_0 . Donc, nous avons utilisé les lois a priori appropriées afin d'éviter que toute loi a posteriori soit inappropriée. Nous pourrions envisager d'utiliser des lois a priori uniformes pour σ^2_i et σ^2_0 , c'est-à-dire $\pi(\sigma^2_i) \propto 1$, et $\pi(\sigma^2_0) \propto 1$, semblables à la loi a priori uniforme sur β . Avec les lois a priori uniformes

présente donc une amélioration par rapport aux estimations directes par sondage.

HB pour les données sur le lait. Comme prévu, sur l'ensemble des 43 régions, le fait de supposer que la variance σ^2_i est connue ou inconnue donne lieu à une variation négligeable des estimations ponctuelles, des erreurs-types et des coefficients de variation, étant donné la grande taille des échantillons pour les 43 régions. Par conséquent, la substitution de $\sigma^2_i = s^2_i$ dans le modèle est raisonnable, lorsque les tailles des échantillons régionaux sont grandes, comme l'illustre clairement cet exemple. En outre, les e.-t. et les c.v. des estimations HB sont plus petits que ceux des estimations directes par sondage présentées au tableau 2. Comme il faut s'y attendre, l'approche HB re-présente donc une amélioration par rapport aux estimations directes par sondage.

Données sur le lait : Le tableau 4 contient les estimations HB pour les données sur le lait. Comme prévu, sur l'ensemble des 43 régions, le fait de supposer que la variance σ^2_i est connue ou inconnue donne lieu à une variation négligeable des estimations ponctuelles, des erreurs-types et des coefficients de variation, étant donné la grande taille des échantillons pour les 43 régions. Par conséquent, la substitution de $\sigma^2_i = s^2_i$ dans le modèle est raisonnable, lorsque les tailles des échantillons régionaux sont grandes, comme l'illustre clairement cet exemple. En outre, les e.-t. et les c.v. des estimations HB sont plus petits que ceux des estimations directes par sondage présentées au tableau 2. Comme il faut s'y attendre, l'approche HB re-présente donc une amélioration par rapport aux estimations directes par sondage.

sont faibles (pour les données sur le maïs, l'erreur-type est de 5,704 et le c.v., de 0,036 pour le comté de Franklin). Comme les tailles d'échantillon sont très faibles, ces erreurs-types d'échantillon ne peuvent être considérées comme des approximations fiables des erreurs-types réelles. Le tableau 1 présente les données de niveau régional modifiées pour le maïs et le soja produites d'après les données au niveau unitaire de Battese et coll. (1988).

Tableau 1

Données de niveau régional modifiées sur les cultures, d'après Battese, Harter et Fuller (1988)

Pays	n_i	y_i	c.v.	e-t.	Soja
Franklin	3	158,623	5,704	0,036	52,473 16,425 0,313
Pocahontas	3	102,523	43,406	0,423	118,697 50,290 0,424
Winnebago	3	112,773	30,547	0,271	88,573 10,453 0,118
Wright	3	144,297	53,999	0,374	97,800 52,034 0,532
Webster	4	117,595	21,298	0,181	112,980 23,531 0,208
Hancock	5	109,382	15,661	0,143	117,478 17,209 0,146
Kossuth	5	110,252	12,112	0,110	117,844 20,954 0,178
Hardin	5	120,054	36,807	0,307	101,834 26,790 0,263

Données sur le lait : Les données sur le lait, utilisées dans un article publié par Arora et Lahiri (1997), proviennent du U.S. Bureau of Labor Statistics. Les valeurs estimées sont les dépenses moyennes en lait frais pour 1989. L'ensemble contient des données sur 43 régions dont la taille d'échantillon varie de 95 à 633. Les c.v. varient de 0,074 à 0,341 sur les 43 régions. Le lecteur trouvera une description plus détaillée des données dans Arora et Lahiri (1997). Par souci de complétude, nous présentons les données au tableau 2. À l'instar d'Arora et Lahiri (1997), nous utilisons $x_i' \beta = \beta_j$ si $i \in j$, grande région (série de régions semblables pour la publication). Arora et Lahiri (1997) ont utilisé huit grandes régions. Puisque cette division en huit grandes régions n'est pas décrite dans leur article, après avoir relevé les tendances dans les données, nous avons utilisé le modèle de Fay-Herriot pour tester deux nouvelles divisions en six et en quatre grandes régions obtenues en regroupant les estimations par sondage semblables. En général, l'utilisation de ces grandes régions produit une réduction importante des c.v. Alors que les six groupes ont produit une réduction moyenne des c.v. d'environ 20 %, les quatre groupes ont donné une réduction moyenne d'environ 25 % des c.v. comparativement aux estimations directes. La comparaison des estimations ponctuelles et des c.v. montre que l'utilisation des quatre grandes régions donne de meilleurs résultats que l'utilisation des six grandes régions. Les quatre grandes régions sont 1-7, 8-14, 15-25 et 26-43. Ici, nous utiliserons ces quatre groupes comme variables auxiliaires aux fins d'illustration.

3.2 Analyse des résultats

Données sur le maïs et le soja : Pour commencer, nous examinons l'effet de notre traitement de σ_i^2 en utilisant l'approche HB. Le tableau 3 donne les estimations HB, $\hat{\theta}_{HB,i}$, et les erreurs-types (e-t.) et les coefficients de variation (c.v.) connexes pour les ensembles de données de niveau régional pour le maïs et le soja. L'erreur-type est la racine carrée de la variance a posteriori. Sous le modèle 1 (σ_i^2 inconnue), les e-t. et les c.v. sont systématiquement plus élevés que les valeurs correspondantes sous le modèle 2 ($\sigma_i^2 = s_i^2$ connue). L'accroissement des e-t. et des c.v. sous le modèle 1 est prévisible, puisque ce modèle tient compte

Tableau 2
Donnée sur le lait, tirées de Arora et Lahiri (1997)

Petite région	n_i	y_i	e-t.	c.v.
1	191	1,099	0,163	0,148
2	633	1,075	0,080	0,074
3	597	1,105	0,083	0,075
4	221	1,028	0,109	0,174
5	195	0,753	0,119	0,158
6	191	0,981	0,141	0,144
7	183	1,257	0,202	0,161
8	188	1,095	0,127	0,116
9	204	1,405	0,168	0,120
10	188	1,356	0,178	0,131
11	149	0,615	0,100	0,163
12	290	1,460	0,201	0,138
13	250	1,338	0,148	0,111
14	194	0,854	0,143	0,167
15	184	1,176	0,149	0,127
16	193	1,111	0,145	0,131
17	218	1,257	0,135	0,107
18	266	1,430	0,172	0,120
19	214	1,278	0,137	0,107
20	213	1,292	0,163	0,126
21	196	1,002	0,125	0,125
22	95	1,183	0,247	0,209
23	195	1,044	0,140	0,134
24	187	1,267	0,171	0,135
25	479	1,193	0,106	0,089
26	230	0,791	0,121	0,153
27	186	0,795	0,121	0,152
28	199	0,759	0,239	0,341
29	238	0,796	0,106	0,133
30	207	0,565	0,089	0,158
31	165	0,886	0,225	0,254
32	153	0,952	0,205	0,215
33	210	0,807	0,119	0,147
34	383	0,582	0,067	0,115
35	255	0,684	0,106	0,155
36	226	0,787	0,126	0,160
37	224	0,440	0,092	0,209
38	212	0,759	0,132	0,174
39	39	0,770	0,100	0,130
40	179	0,800	0,113	0,141
41	312	0,756	0,083	0,110
42	241	0,865	0,121	0,140
43	205	0,640	0,129	0,202

$$\bullet \quad [\sigma_1^2 | y, \theta, \beta, \sigma_1^2] \sim \text{GI} \left(a_1 + \frac{d_1}{1}, b_1, \frac{2}{(y_1 - \theta_1)^2 + d_1 s_1^2} \right) \quad \text{ou } d_1 = n_1 - 1, i = 1, \dots, m;$$

$$\bullet \quad [\sigma_2^2 | y, \theta, \beta, \sigma_2^2] \sim \text{GI} \left(a_0 + \frac{2}{m}, b_0, \frac{2}{\sum_{i=1}^m (\theta_i - x_i^j \beta)^2} \right)$$

Il est facile de tirer des échantillons à partir de ces lois conditionnelles complètes. Pour les applications, nous utilisons $L = 5$ exécutions parallèles, chacune avec une durée de « rodage » de $B = 1\,000$ et une taille d'échantillon de Gibbs de $G = 5\,000$. Les paramètres a priori a_i , b_i et a_0 , b_0 sont fixés à 0,0001. Nous obtenons donc l'estimateur HB de θ_i sous le modèle 1 suivant

$$\hat{\theta}_{\text{HB}} = (LG)^{-1} \sum_{G=1}^L \sum_{g=1}^G (y_{i(g)}^j y_i + (1 - y_{i(g)}^j) x_i^j \beta_{i(g)}), \quad (4)$$

où $y_{i(g)}^j = \sigma_{2(i(g))}^2 / (\sigma_{2(i(g))}^2 + \sigma_{2(i(g))}^2)$, et la variance a posteriori de θ_i peut être estimée par

$$V(\theta_i) = (LG)^{-1} \sum_{G=1}^L \sum_{g=1}^G (y_{i(g)}^j \sigma_{2(i(g))}^2 + (LG)^{-1} \sum_{G=1}^L \sum_{g=1}^G (y_{i(g)}^j y_i + (1 - y_{i(g)}^j) x_i^j \beta_{i(g)})^2$$

$$- \left\{ (LG)^{-1} \sum_{G=1}^L \sum_{g=1}^G (y_{i(g)}^j y_i + (1 - y_{i(g)}^j) x_i^j \beta_{i(g)})^2 \right\}, \quad (5)$$

où $\{\beta_{i(g)}^j, \sigma_{2(i(g))}^2, y_{i(g)}^j, g = 1, \dots, G, i = 1, \dots, L\}$ est

l'échantillon généré au moyen de l'échantillonneur de Gibbs. Les estimateurs (4) et (5) sont les estimateurs HB dits rao-blackwellisés. Les estimateurs rao-blackwellisés sont plus stables pour ce qui est des erreurs de simulation, comme l'ont montré, par exemple, Gelfand et Smith (1991), ainsi que You et Rao (2000).

Maintenant, considérons le modèle 2. Les lois conditionnelles complètes pour l'échantillonneur de Gibbs sous le

$$\bullet \quad [\theta_i | y, \beta, \sigma_2^2] \sim N(y_i, y_i + (1 - y_i)(x_i^j \beta_j, y_i s_i^2)), \quad \text{ou}$$

$$y_i = \frac{\sigma_2^2}{\sigma_2^2 + s_i^2}, \quad i = 1, \dots, m;$$

$$\bullet \quad [\beta | y, \theta, \sigma_2^2] \sim N^d \left(\left(\sum_{i=1}^m x_i^j x_i^j \right)^{-1} \left(\sum_{i=1}^m x_i^j y_i \right), \left(\sum_{i=1}^m x_i^j x_i^j \right)^{-1} \right);$$

Sous le modèle 2, l'estimateur HB de θ_i et l'estimateur de la variance a posteriori correspondant sont donnés par (4) et (5), respectivement, avec $\sigma_{2(i(g))}^2$ remplacé par s_i^2 . Soulignons que l'utilisation de s_i^2 au lieu de $\sigma_{2(i(g))}^2$ peut donner lieu à une sous-estimation importante de la variance a posteriori de θ_i pour certaines régions pour lesquelles la taille d'échantillon n_i est petite. Nous comparerons les estimateurs HB et évaluerons les effets de l'utilisation de s_i^2 dans le modèle 2 grâce à une analyse de données à la section suivante.

3. Analyse de données

3.1 Les ensembles de données

Nous considérons deux ensembles de données intéressants pour nos analyses. Le premier contient des données sur les cultures de maïs et de soja pour huit régions seulement pour lesquelles la taille d'échantillon est petite. Le deuxième contient des données sur le lait pour 43 régions pour lesquelles la taille d'échantillon est relativement grande. Nous comparerons les modèles HB et les estimations basées sur ces deux ensembles de données.

Données sur le maïs et le soja: Ces données, qui proviennent du U.S. Department of Agriculture, ont été étudiées pour la première fois par Batteas, Harter et Fuller (1988). Elles contiennent les nombres d'hectares cultivés déclarés et des données recueillies par le satellite LANDSAT pour les cultures de maïs et de soja dans des segments échantillonnés de 12 comtés de l'Iowa. Les nombres déclarés d'hectares pour chaque culture constituent les estimations directes par sondage. Les données auxiliaires sont les moyennes de population du nombre de pixels d'une culture donnée par segment. Les tailles d'échantillon sont petites pour ces régions, variant de un à cinq. Pour l'étude, nous utilisons uniquement les comtés pour lesquels la taille d'échantillon est égale ou supérieure à trois (huit régions répondent à ce critère). Par conséquent, la taille d'échantillon des comtés varie de trois à cinq. Les données originales sont des données au niveau de l'unité. Afin d'obtenir des données au niveau de la région, nous avons calculé la moyenne d'échantillon et l'erreur-type d'échantillon pour chaque comté. Pour les données sur le maïs et le soja, les erreurs-types d'échantillon sont en général assez grandes (donnant certains c.v. dans la fourchette de 0,3 à 0,4 et un c.v. de 0,532), mais, par hasard, dans certains cas elles

Le modèle bien connu de Fay-Herriot (Fay et Herriot 1979)

Modèle 1

- $y_i | \theta_i, \sigma_i^2 \sim \text{ind } N(\theta_i, \sigma_i^2), i = 1, \dots, m;$
- $d_i s_i^2 | \sigma_i^2 \sim \text{ind } \chi_{d_i}^2, d_i = n_i - 1, i = 1, \dots, m;$
- $\theta_i | \beta, \sigma_i^2 \sim \text{ind } N(x_i' \beta, \sigma_i^2), i = 1, \dots, m;$
- Lois a priori sur les paramètres : $\pi(\beta) \propto 1, \pi(\sigma_i^2) \sim \text{GI}(a_i, b_i), (0 \leq i \leq m)$ sont des constantes connues fixées à une valeur très petite afin de refléter les connaissances vagues au sujet de σ_i^2 et σ_v^2 . GI dénote la loi gamma inverse.

Dans le modèle 1, les variances d'échantillonnage σ_i^2

sont inconnues. Cependant, en pratique, nous pourrions avoir un modèle plus simple en remplaçant σ_i^2 par son estimation s_i^2 (ici s_i^2 est traitée comme si elle était constante) et obtenir le modèle suivant :

Modèle 2

- $y_i | \theta_i \sim \text{ind } N(\theta_i, s_i^2), i = 1, \dots, m;$
- $\theta_i | \beta, \sigma_i^2 \sim \text{ind } N(x_i' \beta, \sigma_i^2), i = 1, \dots, m;$
- Lois a priori : $\pi(\beta) \propto 1, \pi(\sigma_i^2) \sim \text{GI}(a_i, b_i).$

Nous voulons estimer les paramètres de petite région θ_i .

Rivest et Vandal (2002), ainsi que Wang et Fuller (2003) ont obtenu les estimateurs par la méthode empirique du meilleur prédicteur linéaire sans biais (EBLUP) de θ_i et les approximations des erreurs quadratiques moyenne (EQM) associées en supposant que m et n_i sont relativement grands. Dans le présent article, nous considérons une approche hiérarchique bayésienne (HB) s'appuyant sur la méthode d'échantillonnage de Gibbs. L'un des avantages de l'approche HB est qu'elle est simple et que les inférences pour les paramètres θ_i sont « exactes », contrairement à celles obtenues par l'approche EBLUP. Le paramètre de petite région θ_i est estimé par sa moyenne a posteriori et sa précision est mesurée par sa variance a posteriori. L'approche HB tient compte automatiquement des incertitudes associées aux paramètres inconnus dans le modèle. À la section 2, nous présentons les modèles régionaux HB et les inférences basées sur l'échantillonnage de Gibbs connexes. À la section 3, nous décrivons l'analyse de deux ensembles de données d'enquête et une analyse de sensibilité. Enfin, à la section 4, nous offrons certaines conclusions et proposons certaines orientations pour des futurs travaux.

2. Approche hiérarchique bayésienne

Nous allons maintenant présenter le modèle régional (3) et les variances d'échantillonnage estimées s_i^2 dans un cadre hiérarchique bayésien (HB) comme il suit :

$$[\beta | y, \theta, \sigma_i^2, \sigma_v^2] \sim N^p \left(\sum_{i=1}^m x_i' x_i' \theta_i, \sum_{i=1}^m \sigma_i^2 x_i' x_i' \right);$$

$$y_i = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_i^2}, i = 1, \dots, m;$$

$$[\theta_i | y, \beta, \sigma_i^2, \sigma_v^2] \sim N(\gamma_i y_i + (1 - \gamma_i) x_i' \beta, \gamma_i \sigma_i^2), \text{ où}$$

complètes suivantes pour l'échantillonneur de Gibbs :

Sous l'approche HB, nous utilisons la moyenne a posteriori $E(\theta_i | y)$ en tant qu'estimation ponctuelle de θ_i et la variance a posteriori $V(\theta_i | y)$ en tant que mesure de la variabilité, où $y = (y_1, \dots, y_m)'$. Pour estimer $E(\theta_i | y)$ et $V(\theta_i | y)$, nous employons la méthode d'échantillonnage de Gibbs (Gelfand et Smith 1990). Partant du modèle 1, nous obtenons les lois conditionnelles

Estimation pour petits domaines au moyen de modèles d'échantillonnage et d'estimations des variances d'échantillonnage

Yong You et Beatrice Chapman¹

Résumé

Dans le contexte de l'estimation pour petits domaines, des modèles régionaux, comme le modèle de Fay-Herriot (Fay et Herriot, 1979), sont très souvent utilisés en vue d'obtenir de bons estimateurs fondés sur un modèle pour les petits domaines ou petites régions. Il est généralement supposé que les variances d'erreur d'échantillonnage incluses dans le modèle sont communes. Dans le présent article, nous considérons la situation où les variances d'erreur d'échantillonnage sont estimées individuellement au moyen d'estimateurs directs. Nous construisons un modèle hiérarchique bayésien (HB) complet pour les estimateurs par sondage directs et pour les estimateurs de variance de l'erreur d'échantillonnage. Nous employons la méthode d'échantillonnage de Gibbs pour obtenir les estimateurs HB pour les petites régions. L'approche HB proposée tient compte automatiquement de l'incertitude supplémentaire associée à l'estimation des variances d'erreur d'échantillonnage, particulièrement quand la taille des échantillons régionaux est très faible. Nous comparons le modèle HB proposé au modèle de Fay-Herriot grâce à l'analyse de deux ensembles de données d'enquête. Nos résultats montrent que les estimateurs HB proposés donnent d'assez bons résultats comparativement aux estimations directes. Nous discutons également du problème des lois a priori sur les composantes de la variance.

Mots clés : Échantillonnage de Gibbs; hiérarchique bayésien; sensibilité aux lois a priori; taille d'échantillon; composantes de la variance.

1. Introduction

Dans la plupart des applications, les enquêtes par sondage sont conçues afin de fournir des estimations directes fiables pour l'ensemble de la population, de même que pour les grandes régions au moyen de données d'échantillon propres à la région. Toutefois, fréquemment, cette méthode d'estimation directe ne produit pas d'estimations fiables pour les petites régions, à cause de la très petite taille des échantillons obtenus pour ces dernières. Puisque les estimations directes pour les petites régions sont souvent assorties d'une erreur-type trop grande, si l'on veut augmenter la précision et la fiabilité, il est nécessaire d'« emprunter de l'information » aux régions apparentées, donc d'accroître la taille efficace de l'échantillon, en vue de produire des estimations indirectes pour les petites régions (Rao 1999). Les méthodes fondées sur un modèle explicite, qui s'appuient sur des données supplémentaires, telles que des données de recensement ou des données administratives, associées aux petites régions dans des modèles explicites en vue de relier ces régions, ont été utilisées à grande échelle en pratique pour obtenir des estimateurs fondés sur un modèle fiables. Ces modèles se répartissent en deux grandes catégories, à savoir les modèles au niveau de la région et les modèles au niveau de l'unité. Les modèles de niveau régional sont fondés sur des estimateurs par sondage régionaux directs et les modèles de niveau unitaire sont fondés sur les observations individuelles recueillies dans les régions. Pour une vue d'ensemble et une évaluation des modèles appliqués à l'estimation pour petits domaines

ou petites régions, voir Rao (1999, 2003). Dans le présent article, nous étudions les modèles de niveau régional.

Pour obtenir un modèle régional de base, nous supposons que le paramètre d'intérêt de la petite région θ_i est relié à des données auxiliaires propres à la région $x_i = (x_{i1}, \dots, x_{ip})$ grâce à un modèle linéaire

$$(1) \quad \theta_i = x_i' \beta + v_i, \quad i = 1, \dots, m,$$

où m est le nombre de petites régions, $\beta = (\beta_1, \dots, \beta_p)'$ est le vecteur de dimensions $p \times 1$ de coefficients de régression, et les v_i sont les effets aléatoires propres à la région que nous supposons être indépendants et identiquement distribués (iid) avec $E(v_i) = 0$ et $\text{var}(v_i) = \sigma_v^2$. L'hypothèse de normalité peut également être incluse. Ce modèle est appelé modèle de liaison pour θ_i .

Le modèle régional de base repose aussi sur l'hypothèse qu'étant donné la taille d'échantillon propre à la région $n_i > 1$, il existe un estimateur par sondage direct y_i (habituellement sans biais par rapport au plan de sondage) pour le paramètre de petite région θ_i , tel que

$$(2) \quad y_i = \theta_i + e_i, \quad i = 1, \dots, m,$$

où e_i est l'erreur d'échantillonnage associée à l'estimateur direct y_i . Nous supposons aussi que les e_i sont des variables aléatoires normales indépendantes de moyenne $E(e_i | \theta_i) = 0$ et de variance d'échantillonnage $\text{var}(e_i | \theta_i) = \sigma_e^2$. La combinaison des modèles (1) et (2) donne un modèle linéaire mixte régional

$$(3) \quad y_i = x_i' \beta + v_i + e_i, \quad i = 1, \dots, m.$$

perte plus importante de précision qu'un accorèissement unitaire. Par consèquent, les plans dans lesquels la variance d'échantillonnage (estimée par rééchantillonnage) des tailles des sous-échantillons $n_p(d)$ fixè est plus faible sont mieux adaptés à l'estimation pour petits domaines. Dans les plans adaptés à l'estimation pour petits domaines, l'échantillon domaine pourrait ne pas ètre représentè dans l'échantillon lors de certaines répliques et pourrait ètre surreprésentè plusieurs fois dans d'autres. En gènéral, il est préférable d'utiliser de plus petites grappes pour l'estimation pour petits domaines, si cela n'augmente pas les coûts d'enquète et qu'il est possible de maintenir une taille globale d'échantillon fixe.

Remerciements

Je remercie le rédacteur en chef dèlèguè et les examinateurs d'avoir proposè plusieurs améliorations, mais surtout de m'avoir fait dècouvrir une erreur dans une version antérieure du manuscrit. Je tiens aussi à mentionner mes discussions avec l'équipe polonaise du projet EURAREA.

Bibliographie

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Goldstein, H. (1995). *Multilevel Statistical Models*. Deuxième Edition. Edward Arnold, London, UK.

Longford, N.T. (1993). *Random Coefficient Models*. Oxford University Press, Oxford.

Longford, N.T. (1999). Multivariate shrinkage estimation of small-area means and proportions. *Journal of the Royal Statistical Society, Series A*, 162, 227-245.

Longford, N.T. (2004). Missing data and small area estimation in the UK Labour Force Survey. *Journal of the Royal Statistical Society, Series A*, 167, 341-373.

Longford, N.T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. Springer-Verlag, New York.

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

Marker, D.A. (2001). Production d'estimations régionales d'après les données d'enquêtes nationales : Méthodes visant à réduire au minimum l'emploi d'estimateurs indirects. *Techniques d'enquête*, 27, 201-207.

Platek, R., Rao, J.N.K., Samdal, C.-E. et Singh, M.P. (Eds.) (1987). *Small Area Statistics*. New York: John Wiley & Sons.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Singh, M.P., Gambino, J. et Mantel, H.J. (1994). Les petites régions : Problèmes et solutions. *Techniques d'enquête*, 20, 3-23.

simples, pourrait ètre plus pratique. En outre, les priorités, ou l'opinion d'experts à leur sujet, peuvent èvoluer au cours du temps, même pendant la réalisation de l'enquète et l'analyse des données. Les estimations associées à une erreur-type ou à un coefficient de variation supérieur à un seuil précisè sont souvent exclues des rapports analytiques. L'intention de les exclure peut ètre reflétèe dans le calcul de la taille d'échantillon en considérant θ comme étant l'estimateur de θ_p , c'est-à-dire en fixant l'EQM comme à $l'EQM \sigma_p^2 + var(\theta)$ correspondante ou à une autre (grande) valeur constante.

Bien que nous disposions une classe particulière de priorités pour les petits domaines, aucune difficulté conceptuelle ne se pose lorsque l'on utilise une autre classe. Elle pourrait dépendre de plusieurs grands facteurs de population plutôt que de la taille de population uniquement. En principe, les priorités peuvent aussi ètre fixées individuellement pour les petits domaines, bien que cela ne soit pratique que si leur nombre est faible. Les priorités fondées sur la formule ou établies individuellement peuvent ètre combinées en ajustant les priorités, telles que $P_p = N_p^d$, pour quelques petits domaines afin de reflèter leur rôle exceptionnel dans l'analyse.

Une analyse de sensibilité, en vue d'étudier les modifications du plan d'échantillonnage en fonction de diverses données d'entrée est essentielle à la compréhension de l'incertitude au sujet des paramètres estimés (le ratio de variance ou en particulier) et le caractère arbitraire, aussi limité qu'il soit, de l'établissement des priorités. Pour cela, il est préférable de disposer d'une solution analytiquement simple qui peut ètre exécutée de nombreuses fois, pour une gamme de conditions, plutôt qu'une solution plus complexe, dont les propriétés sont difficiles à étudier.

Les estimateurs composites multivariés exploitent la similarité non seulement entre les petits domaines, mais aussi entre les variables (auxiliaires), les périodes, les sous-populations, et ainsi de suite (Longford 1999 et 2005). L'EQM de ces estimateurs dépend de la manière de variantes mise à l'échelle Ω , qui est le pendant multivarié de ω . Le calcul de la taille d'échantillon par cette méthode est difficile à appliquer directement, parce que, dans Ω , les variances et les covariances sont les unes et les autres essentielles à l'efficacité des estimateurs. Une approche plus constructive consiste à faire concorder la manière Ω avec un ratio ω qui peut ètre interprétè comme étant la similarité des petits domaines après correction pour l'information auxiliaire, comme dans les méthodes bayésiennes empiriques.

Lorsque il est impossible d'exercer un contrôle sur les tailles d'échantillon affectées aux petits domaines, leur calcul demeure utile comme indication de la façon dont elles devraient ètre réparties *en moyenne*. En gènéral, une réduction unitaire de la taille d'échantillon est associée à une

L'approche montre que les tailles d'échantillon optimales sont presque constantes dans la fourchette $\omega \in (\omega_*, +\infty)$, ω_* augmente avec q , G et $1/n$. Il s'agit d'une conséquence de la taille d'échantillon relativement grande n , qui assure que les sous-échantillons de la part importante d'information entre les cantons aient lieu, à moins que les cantons soient fort semblables ($\omega < \omega_*$). La plupart des coefficients de rétrécissement $b_j = 1/(1+n_j\omega)$ sont très petits. Lorsqu'une taille $n = 10\,000$ est prévue, pour les valeurs faibles de ω , la taille d'échantillon optimale augmente fortement pour les cantons les moins peuplés et chute brusquement pour les plus peuplés. La dispersion des tailles d'échantillon optimales augmente avec G et G , convergeant vers la répartition optimale pour l'estimation de la moyenne nationale θ , qui correspond à $\omega = 0$. Par contre, les tailles d'échantillon optimales sont discontinues à $\omega = 0$ quand $G = 0$; les solutions divergent vers $-\infty$ pour les cantons les moins peuplés.

Dans les volets C et D, pour $n = 1\,000$, la variation des tailles d'échantillon en fonction de ω persiste pour une plus grande fourchette de valeur de ω , parce que la portée de l'échantillon d'information entre les cantons est plus grande pour les tailles d'échantillon plus petites. Les tailles d'échantillon optimales ne sont pas des fonctions monotones de ω ; pour les cantons les moins peuplés, on observe un creux pour les faibles valeurs de ω . Le creux est plus prononcé pour les faibles valeurs de G et pour les grandes valeurs de q , c'est-à-dire lorsque les disparités entre les valeurs de q , c'est-à-dire lorsque les disparités entre les priorités des cantons sont grandes et que l'importance relative de l'inférence au sujet de la moyenne nationale est faible au cas discuté pour $G = 0$. À cause des différences de priorité P_d , une faible réduction de l'EQM pour un canton plus peuplé est préférable à une réduction plus importante pour un canton moins peuplé. Le creux existe aussi quand $n = 10\,000$, mais il est si peu profond et si étroit qu'il n'est pas visible dans les conditions de résolution du graphique. Notons que, dans les volets C et D, l'axe des abscisses possède une fourchette de valeurs de ω trois fois plus grande que dans les volets A et B.

Dans le contexte de l'enquête planifiée, il a été convenu qu'il était peu probable que la valeur de ω soit inférieure à 0,05. Par conséquent, le calcul des tailles d'échantillon a pu être fondé sur l'estimateur direct.

4. Discussion

La méthode décrite dans le présent article permet de déterminer le plan d'échantillonnage optimal pour les conditions artificielles d'échantillonnage stratifié avec

Bien que la solution numérique du problème pour l'estimation composite avec une priorité positive G soit simple et ne présente aucun problème de convergence, il est avantageux de disposer d'une solution analytique, afin de pouvoir étudier une gamme de scénarios. La proximité des solutions obtenues pour les estimations directe et composite donne à penser que la répartition optimale pour l'estimation directe pourrait également s'approcher de la situation optimale pour l'estimation composite avec des valeurs raisonnables de ω , disons, $\omega > 0,05$.

Diverses contraintes de gestion et d'organisation consistent en un autre obstacle à l'application littérale d'un plan d'échantillonnage établi analytiquement. Dans les enquêtes-ménages, il est souvent préférable d'attribuer un quota (presque) complet d'adresses à chaque intervieweur, si bien que l'on accorde la préférence aux tailles d'échantillon qui sont des multiples du quota. Ces considérations et de nombreuses autres contraintes peuvent être intégrées dans le problème d'optimisation, quoiqu'elles soient souvent difficiles à quantifier ou que le concepteur de l'enquête ne soit pas conscient de leur existence à cause d'une communauté imparfaite. L'improvisation, après l'obtention d'un plan d'échantillonnage optimal pour des conditions plus

critique lorsque l'un des cantons n est pas représenté dans l'échantillon. Cette déficience de (3) peut être compensée en fixant la priorité relative G à une valeur positive.

La figure 5 résume l'effet de la priorité relative G et de l'exposant de priorité q sur les tailles d'échantillon optimales pour les cantons le moins et le plus peuplés, ainsi que le canton de Thurgau qui possède la 13^e taille de population par ordre décroissant (médiane), soit 228 000. Chaque valeur de q , indiquée dans le titre, et de G , indiquée en utilisant différents types de lignes, est

$n = 1\,000$.

représentée pour un canton par un graphique de la taille d'échantillon optimale en fonction du ratio de variance ω . La limite de cette fonction lorsque $\omega \rightarrow +\infty$, égale à la taille d'échantillon optimale pour l'estimation directe, est marquée par une barre dans la marge de droite du volet en question. Pour $\omega = 0$, on obtient le plan d'échantillonnage optimal pour l'estimation de la moyenne nationale θ . Les volets A et B au haut de la figure correspondent à la taille d'échantillon globale $n = 10\,000$ et les volets C et D, à

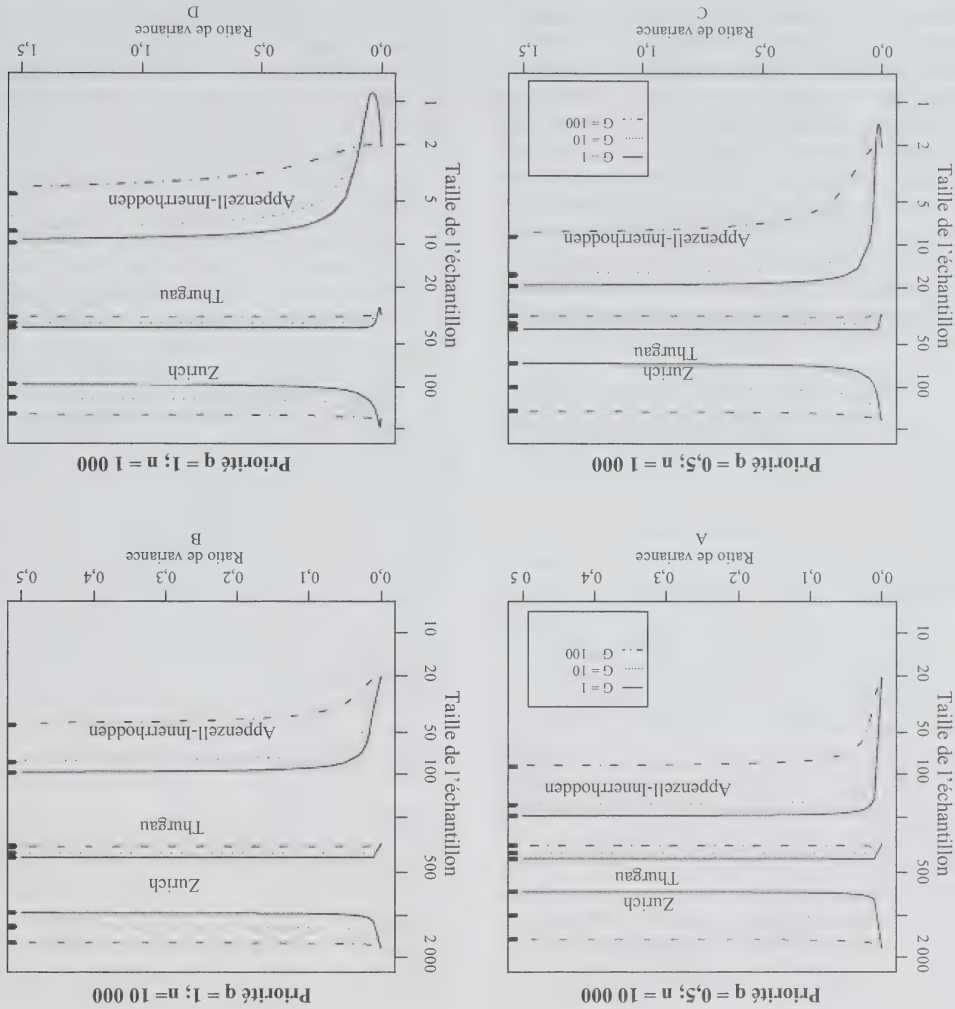


Figure 5. Tailles d'échantillon optimales pour l'estimation composite des moyennes de population pour trois cantons pour une gamme de rapports de variance ω , les exposants de priorité $q = 0,5$ et $q = 1,0$ et les priorités relatives $G = 1, 10$ et 100 . Les tailles globales d'échantillon sont 10 000 (volets A et B) et 1 000 (volets C et D).

$$\sum_{d=1}^D P_d \text{EQM}(\theta_d) + GP^+ v.$$

La solution satisfait la contrainte

$$N_d^p \sigma_d^2 \omega_d^p + GP^+ \frac{N_d}{N_d^p} n_d^p = \text{const.} \quad (4)$$

Cette équation ne possède pas de solution analytique commode, mais elle peut être résolue par application de scénarios itératifs. La valeur de n_d détermine les autres tailles d'échantillon n_d^p , de sorte que l'optimisation correspond à une recherche unidimensionnelle. Si les tailles d'échantillon provisoires n fondées sur un ensemble de valeur de n_d sont trop grandes, on réduit n $\mathbf{1}^T p > n$, n_d et on calcule les autres tailles d'échantillon n_d^p en résolvant (4). Notons que la solution dépend des variances σ_d^2 et σ_d^2 . Le problème se simplifie quelque peu lorsque la variance est la même pour tous les petits domaines $\sigma_d^2 = \sigma_d^2 = \dots = \sigma_d^2$. Alors, la solution de (4) dépend des variances uniquement par la voie du ratio $\omega = \sigma_d^2 / \sigma_d^2$, parce que σ_d^2 est un facteur multiplicatif qui n'a aucun effet sur l'optimisation.

À titre d'exemple, supposons que $q = 1$ et $G = 10$ lors de la planification d'une enquête auprès de la population suisse, avec $n = 10\,000$, et en supposant que $\omega = 0,10$. Comme solution initiale, nous utilisons la répartition optimale pour l'estimation directe avec les mêmes valeurs de q et de G . Une itération met à jour la taille de l'échantillon de chaque canton et, dans les cantons, la mise à jour pour tous, sauf celui de référence sélectionné arbitrairement $d = 1$, est également itérative. La taille provisoire du sous-échantillon pour le canton de référence détermine la valeur courante de la constante dans le deuxième membre de (4). L'équation (4) est résolue, itérativement, pour chaque canton $d = 2, \dots, D$, en utilisant la méthode de Newton. Dans l'application, le nombre d'itérations était inférieur à dix pour chaque canton. Enfin, la taille du sous-échantillon pour le canton de référence est ajustée par le facteur $1/D$ un multiple de la différence entre le total courant des tailles des sous-échantillons et le total cible n . La mise à jour des tailles d'échantillon des cantons est elle-même itérée, mais quelques itérations seulement sont nécessaires pour atteindre la convergence; par exemple, toutes les variations des tailles des sous-échantillons étaient inférieures à 1,0 après trois itérations et inférieures à 0,01 après huit itérations. La convergence est rapide, parce que la solution de départ est proche de la solution optimale; l'écart le plus important entre les deux tailles de sous-échantillon est celui observé pour Zurich, soit 20,0 (de 1 199,5 au départ à 1219,5 après huit itérations). Pour Appenzel-Innerrhodan, la taille d'échantillon est réduite de 81,6 à 73,4. Des changements de moins d'une unité ont lieu pour cinq cantons dont la taille de population varie de 228 000 à 278 000. Notons qu'en

Pas de priorité accordée à l'estimation nationale

Pratique, les tailles des sous-échantillons seraient arrondies et éventuellement ajustées davantage afin de satisfaire aux diverses contraintes de gestion de l'enquête.

Si l'estimation nationale n'a aucune priorité, $G = 0$, l'équation (4) possède la solution explicite

$$n_d^* = \frac{\omega}{n\omega + D} \frac{U^{(b)}_d}{N_d^p} - \frac{\omega}{1},$$

où $U^{(b)}_d = N_d^{1/2} + \dots + N_d^{D/2}$. Cette répartition est reliée à la répartition n_d^* , $d = 1, \dots, D$, qui est optimale pour l'estimation directe de θ_d , par l'identité

$$n_d^* = n_d^1 + \frac{\omega}{1} \left(\frac{U^{(b)}_d}{DN_d^p} - 1 \right).$$

Donc, quand $q > 0$, la répartition optimale est plus dispersée dans le cas de l'estimation composite que dans celui de l'estimation directe, et elle est plus grande pour les petits domaines dont la population $N_d > N_d^1$ est plus petite dans le cas de l'échantillon pour les petits domaines ayant une taille de l'équilibre est $N_d^1 = (U^{(b)}_d / D)^{2/q}$; la taille du sous-celui de l'estimation directe. La taille de population au point dispersé dans le cas de l'estimation composite est plus

Si $\omega = 0$, les équations pour le plan d'échantillonnage optimal donnent lieu à une singularité. Dans ce cas, chaque θ_d est estimé efficacement par l'estimateur national $\hat{\theta}$, si bien que le plan optimal pour l'estimation composite coïncide avec le plan optimal pour l'estimateur national ($n_d^* = nN_d^p / N$). Pour $q > 0$, la répartition optimale donne des tailles d'échantillon négatives n_d^* quand

$$N_d^p > \frac{\left\{ \frac{U^{(b)}_d}{2/q} + D \right\}}{\omega + D}. \quad (5)$$

Cette solution (formelle) n'a pas de sens. Une solution négative ne devrait pas être étonnante, car l'EQM de (3) est une fonction analytique pour $n_d > -1/\omega_d$. Si les valeurs de $\omega > 0$ sont faibles, l'EQM est une fonction décroissante à pente faible de la taille d'échantillon n_d . Une valeur négative de n_d^* indique qu'un « petit » canton ne vaut pas la peine d'être échantillonné, à cause de la faible priorité d'inférence P_d . Bien que l'accroissement de la taille de l'échantillon d'un canton plus peuplé d' puisse donner lieu à une réduction plus faible de l'EQM que cela ne serait le cas pour un petit canton d , la priorité plus grande $P_{d'}$ augmente l'effet.

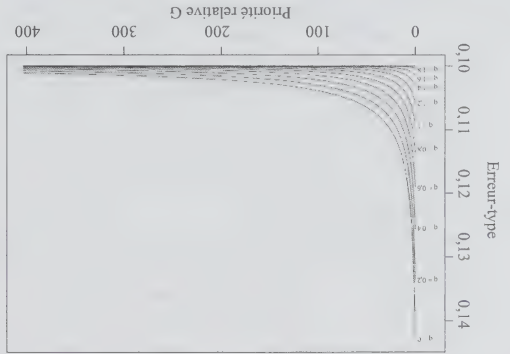
Priorité positive pour la moyenne nationale

Dans (3), l'EQM ne tient pas compte de l'incertitude au sujet de la moyenne nationale θ , situation qui devient

L'utilisation la plus efficace des ressources disponibles pour réaliser une enquête s'obtient par combinaison optimale d'un plan d'échantillonnage et d'un ou de plusieurs estimateurs, si bien que le plan d'échantillonnage et (le choix de) l'estimateur devraient, dans des circonstances idéales, être optimisés simultanément. Ce problème est difficile à résoudre formellement dans la plupart des conditions, quoique certains estimateurs soient plus efficaces que leurs concurrents et que l'on considère une grande gamme de plans d'échantillonnage. Les estimateurs composites (Longford 1999, 2004) représentent l'une de ces classes

3. Estimation composite

Figure 4. Erreur-type de l'estimateur national pour la répartition optimale sous une matrice de priorités données par q et G .



Dans le cas de chaque exposant $q < 2$, la courbe de répartition de la taille de l'échantillon $n_q(G)$ montre une diminution pour les cantons les moins peuplés et une augmentation pour les plus peuplés en direction de la représentation proportionnelle, $n_p = n^{N_p}/N$, qui correspond à $q = 2$. Sur l'échelle linéaire, l'augmentation est assez rapide pour Zürich pour les faibles valeurs de q et de G , tandis que la réduction pour Appenzell-Innerehoden est plus progressive. À mesure que la priorité relative G est réduite, la taille d'échantillon excédentaire est réduite de Zürich (et de quelques autres cantons peuplés) à plusieurs cantons moins peuplés.

La figure 4 représente graphiquement l'erreur-type « nationale » $\sqrt{\text{var}(\theta)}$ sous la répartition optimale de l'échantillon pour une matrice de valeurs de q et de G . Le graphique montre qu'une légère augmentation de G aux alentours de $G = 0$ réduit spectaculairement l'erreur-type de θ , tandis que pour les valeurs plus grandes de G , l'erreur-type ne varie que légèrement. Pour chaque G , un exposant de priorité plus élevé q est associé à une précision plus élevée de θ .

l'estimation pour petits domaines. Pour les estimateurs composites des moyennes de petit domaine, nous recherchons la répartition de l'échantillon qui minimise la fonction objectif

où « aEOM » dénote l'EOM dans laquelle Δ_d^2 est remplacé par σ_d^2 , sa moyenne sur l'ensemble des petits domaines. Dans (3), aEOM est aussi une approximation de la variance conditionnelle de l'estimateur EBUP de la moyenne au niveau du petit domaine fondée sur le modèle (empirique bayésien) à deux niveaux (Longford 1993, Goldstein 1995, Markier 1999 et Rao 2003). Voir Ghosh et Rao (1994) pour une revue reconnue de l'application de ces modèles à

$$(3) \quad \sigma_B^2(\tilde{\theta}^p) = \frac{1 + u^p \omega^p}{\sigma_B^2},$$

précise que ne le sont la plupart des Δ^d_p . Si les coefficients b_p sont estimés avec suffisamment de précision, l'estimateur composite θ_p est plus efficace que les deux estimateurs qui le constituent, θ_p et θ . Si nous ne tenons pas compte de l'incertitude au sujet des variances intra et interdomaines, ni au sujet de la moyenne nationale $\bar{\theta}$ et de la corrélation entre les estimateurs (direct) au niveau national et sur petits domaines, l'EQM moyenne de $\bar{\theta}_p$ est

Si les écarts $\Delta^p = \theta - \underline{\theta}$ étaient connus, le coefficient optimal b^* dans (2) serait, approximativement, $b^* = \sigma^2 / (\sigma^2 + m^p \Delta^p)$. Puisque nous ne connaissons pas Δ^p (sinon, nous serions estimés avec une grande précision par $\theta + \Delta^p$), nous remplaçons Δ^p par sa moyenne sur les petits domaines, égale à σ^p , ce qui donne le coefficient $b = 1 / (1 + m^p \omega)$, où $\omega = \sigma^2 / \sigma^p$ est le ratio de variance. La variance σ^2 doit aussi être estimée, mais, si le nombre de petits domaines est élevé, l'estimation est beaucoup plus

avec des coefficients particuliers aux petits domaines $\hat{\theta}_j$ qui sont de la similarité de l'optimum. La composition θ_j tire parti de la similarité des petits domaines et est particulièrement efficace lorsqu'ils présentent une faible variance interdomaines $\sigma^2_B = D^{-1} \Sigma^{-1} (\theta_j - \bar{\theta})^2$, où $\bar{\theta} = D^{-1} \Sigma^{-1} \theta_j$. Cette variance est définie sur les D paramètres de la population θ_j et n'est pas affectée par le plan d'échantillonnage. En pratique, il faut estimer σ^2_B . Lors de la planification d'une enquête, il est nécessaire d'utiliser des estimations provenant d'autres enquêtes, et de tenir compte de l'incertitude au sujet de σ^2_B , ce qui peut se faire par une analyse de sensibilité, en recherchant les plans d'échantillonnage optimaux pour une gamme de valeurs plausibles de

$$(7) \quad \theta^p q + {}^p \theta ({}^p q - 1) = {}^p \theta$$

d'estimateurs. Il s'agit de combinaisons convexes des estimateurs directs sur petits domaines et au niveau national,

Cette solution correspond à un ajustement des priorités P_d par $GP^+N_d^2/N_2$. Notons que cet ajustement n'est ni additif, ni multiplicatif. L'accroissement de la priorité est plus important pour les petits domaines plus peuplés. Par conséquent, les tailles des sous-échantillons de petit domaine sont réduites davantage quand la priorité relative de l'estimation nationale est intégrée et que les priorités au niveau des petits domaines ne changent pas. La correction pour population finie n'a aucun effet sur n_d^* , parce qu'elle réduit chaque variance d'échantillonnage v_d et v d'une quantité qui ne dépend pas de n .

La priorité G peut être fixée en insistant sur le fait que la perte d'efficacité lors de l'estimation de la grandeur nationale θ n'excède pas un pourcentage donné ou qu'au plus, quelques-uns seulement des écarts absolus $|P_d' - P_d|$ ou des logarithmes des ratios $|\log(P_d'/P_d)|$ (voire aucun) ne soient très grands. Cependant, le problème analytique est facile à résoudre, de sorte que la gestion de l'enquête peut être présentée au moyen des plans d'échantillonnage qui sont optimaux pour une gamme de valeur G .

La variation de la taille des sous-échantillons en fonction de l'exposant q et de la priorité relative G est représentée graphiquement à la figure 3 pour les cantons le moins et le plus peuplés, Appenzell-Innerrhoden et Zurich, dans les volets A et C. Les volets B et D donnent la représentation des mêmes courbes qu'A et C, respectivement, sur l'échelle logarithmique. Ne pas tenir compte de l'objectif de production d'une estimation nationale correspond au cas où $G = 0$ et ne pas tenir compte de l'objectif de production d'une estimation pour petits domaines correspond au cas des valeurs très grandes de G . Tout au long de l'article, nous supposons que $n = 10\,000$ et que $\sigma^2 = 100$ pour tous les cantons.

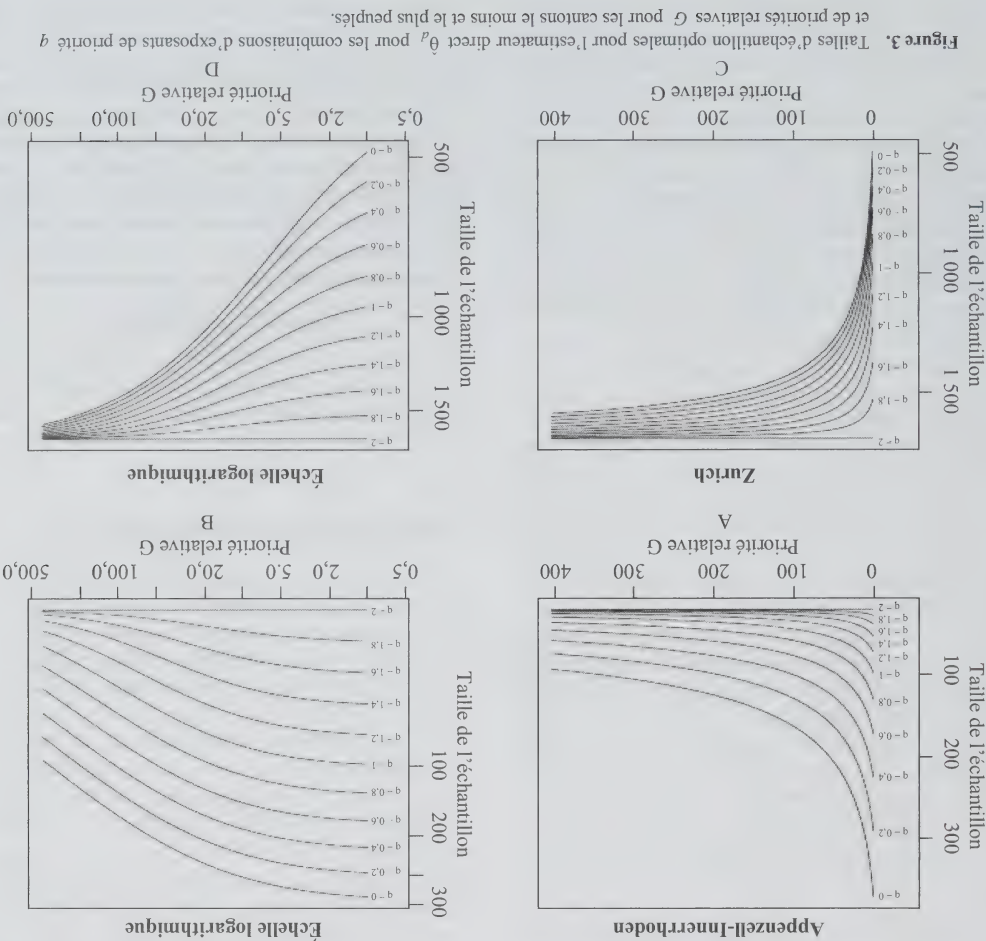


Figure 3. Tailles d'échantillon optimales pour l'estimateur direct θ_d pour les combinaisons d'exposants de priorité q et de priorités relatives G pour les cantons le moins et le plus peuplés.

Pour $q = 0$, une même taille d'échantillon est attribuée à chaque canton, soit $10\,000/26 = 385$, et pour $q = 2$, la répartition est proportionnelle à la taille de population du canton. Pour les valeurs intermédiaires de q , les tailles d'échantillon des cantons les moins peuplés sont augmentées par rapport à la répartition proportionnelle ($q = 2$), au prix de l'attribution d'une taille réduite aux cantons les plus peuplés. Pour les cantons dont la population est supérieure à 250 000, environ 3 % du chiffre national de population, la taille des sous-échantillons dépend fort peu de la valeur de q .

2.1 Priorité accordée à l'estimation nationale

Comme les tailles de sous-échantillon au niveau du canton est assortie d'une perte d'efficacité de l'estimateur tant de priorité $q < 2$, l'estimation optimale au niveau du canton est assortie d'une perte d'efficacité de l'estimateur national. Considérons l'estimateur stratifié

$$\hat{\theta} = \frac{1}{D} \sum_{d=1}^D N_d \hat{\theta}_d$$

de la moyenne nationale θ d'une variable, où $\hat{\theta}_d$ représente les estimateurs sans biais des moyennes intracanton de la même variable. En supposant que l'échantillonnage est stratifié avec échantillonnage aléatoire simple dans les strates (cantons) et que la valeur de θ_d est fixée à la moyenne d'échantillon intrastate,

$$\text{var}(\hat{\theta}) = \frac{1}{D} \sum_{d=1}^D \frac{N_d^2}{N^2} (1 - f_d) \sigma_d^2,$$

où $f_d = n_d / N_d$ est la correction pour population finie. La figure 2 représente la fonction qui relie l'erreur-type $\sqrt{\text{var}(\hat{\theta})}$ à l'exposant de priorité q , calculée en supposant que $\sigma^2 = 100$. L'erreur-type est une fonction décroissante de q , elle diminue plus rapidement à $q = 0$ qu'à $q = 2$, où elle est relativement constante. Pour $q = 2$, les objectifs d'estimation au niveau du canton et au niveau national concordent, et $\sqrt{\text{var}(\hat{\theta})} = 0,100$. Pour $q = 0$, $\sqrt{\text{var}(\hat{\theta})} = 0,143$; dans ces conditions, l'optimalité de l'estimation pour petits domaines a sur l'estimation nationale un effet défavorable important, équivalant à la réduction de moitié de la taille de l'échantillon ($0,143/0,100 = \sqrt{2}$). Pour une valeur négative de q , cet effet est encore plus prononcé.

Donc, nous pouvons répondre au besoin d'efficacité de l'estimateur national en augmentant la valeur de l'exposant de priorité. Par exemple, les parties ayant des intérêts concurrents en matière d'inférence pourraient négocier la perte d'efficacité de $\hat{\theta}$ qu'elles jugent acceptables et fixer ensuite l'exposant de priorité de façon à évaluer cette perte. Ou bien, la perte pourrait être prise en considération lors de l'application du plan d'échantillonnage optimal pour

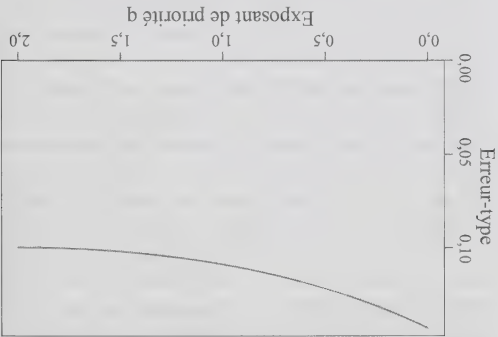


Figure 2. Erreur-type de l'estimateur national $\hat{\theta}$ de la moyenne d'une variable, sous forme de fonction de l'exposant q pour les priorités de l'estimation au niveau du canton.

Un aspect insatisfaisant de ces approches est qu'elles compromettent l'objectif premier des priorités P_d , c'est-à-dire refléter l'importance relative des inférences au sujet de petits domaines distincts. Pour contourner cet inconvénient, nous associons $\hat{\theta}$ à une priorité, dénotée G , relative à une estimation optimale de l'ensemble de D paramètres cibles au niveau du petit domaine θ_d en même temps que le paramètre cible nationale θ . Donc, nous minimisons la fonction objectif

$$\sum_{d=1}^D P_d^v (n_d) + GP^v(\mathbf{n}),$$

où $v = \text{var}(\hat{\theta})$ et $P_d = \mathbf{P}^T \mathbf{1}_D$. Le facteur P_d^+ est introduit pour améliorer l'effet des tailles absolues de P_d et du nombre de petits domaines sur la priorité relative G . Les priorités P_d peuvent être interprétées uniquement d'après leurs tailles relatives, car, pour toute constante $c > 0$, P_d et cP_d correspondent à des ensembles identiques de priorités pour l'estimation pour petits domaines dans (1). Lorsque le plan d'échantillonnage dans chaque petit domaine est aléatoire simple et que $\hat{\theta}$ est l'estimateur stratifié standard, le minimum est atteint quand

$$\sigma_2^d \frac{P_d^+}{P_d^+} = \text{const},$$

où $P_d^+ = P_d + GP_d^+ / N^2$. Les tailles optimales d'échantillon pour les petits domaines sont

$$n_d^* = n \frac{\sigma_1 \sqrt{P_d^+} + \dots + \sigma_D \sqrt{P_d^+}}{\sigma_1 \sqrt{P_d^+}}.$$

que q augmente, une importance relativement plus grande est accordée aux petits domaines les plus peuplés. Lorsque l'échantillon pour $q = 2$, $n_p^1 = n N^p / N$ est proportionnelle aux tailles de population dans les petits domaines et le même plan d'échantillonnage est donc optimal pour les inférences calculées au niveau national et du petit domaine. Pour $q > 2$, la répartition de la taille de l'échantillon est encore plus généreuse à l'égard des petits domaines les plus peuplés, aux dépens de ceux qui le sont moins. Comme cette situation est contre-intuitive dans le contexte de l'estimation pour petits domaines, le choix d'un exposant $q > 2$ n'est probablement jamais approprié. Un exposant de priorité q négatif conviendrait pour une enquête dont le but est de se concentrer sur les petits domaines les moins peuplés. Naturellement, ce genre de plan est très inefficace pour l'estimation du paramètre θ de niveau national, surtout si les tailles de population des petits domaines sont très dispersées.

Les priorités d'inférence P_d peuvent être des fonctions d'autres paramètres que N_d . Par exemple, les tailles de certaines sous-population présentant un intérêt particulier, comme une minorité ethnique dans le petit domaine, peuvent être utilisées au lieu de N_d . P_d peut être défini différemment dans les diverses régions du pays, ou bien la formule pour le calculer peut-être outrepassée pour un petit domaine ou quelques-uns d'entre eux.

Dans certains rapports d'analyse de données d'enquête, une estimation n'est publiée que si elle est fondée sur un

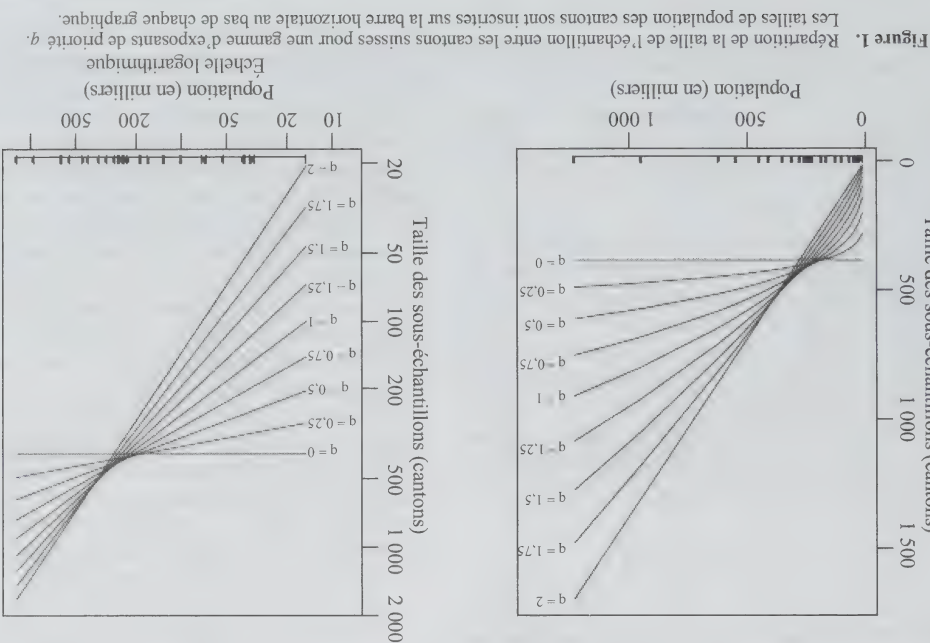
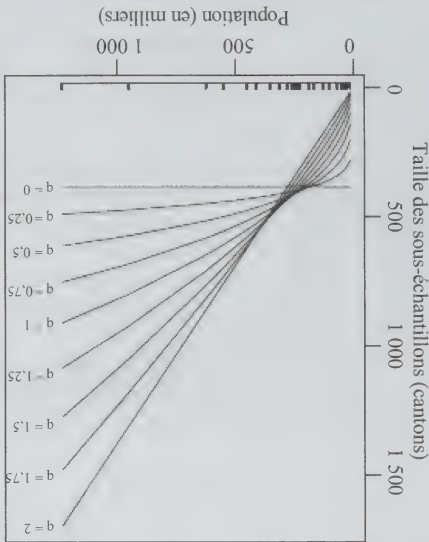


Figure 1. Répartition de la taille de l'échantillon entre les cantons suisses pour une gamme d'exposants de priorité q . Les tailles de population des cantons sont inscrites sur la barre horizontale au bas de chaque graphique.

échantillon de taille suffisamment grande ou que son coefficient de variation (le ratio de l'erreur-type estimée à l'estimation) est inférieur à un seuil spécifié. Si une « pénalité » associée au fait de ne pas publier un paramètre est pré-cisée, elle peut être intégrée dans la définition des priorités fonction objectif (1) soit discontinue et que l'on ne puisse plus appliquer les approches standard d'optimisation. La pénalité doit être déterminée minutieusement. Si elle est trop faible, elle est inefficace; si elle est trop élevée, la solution accordée la préférence à la publication d'estimations pour un aussi grand nombre de petits domaines que possible, mais avec, pour chacun, une taille d'échantillon ou une précision qui n'exécède que de justesse le seuil fixé. Voir Marker (2001) pour une autre approche de ce problème.

La figure 1 illustre l'effet de l'exposant de priorité q sur la répartition de la taille de l'échantillon d'une enquête planifiée en Suisse dans le but d'estimer les moyennes de population d'une variable dans les 26 cantons, en supposant qu'ils ont tous la même variance intracanton σ^2 . La taille globale prévue de l'échantillon est $n = 10\,000$. Dans n'importe quel volet, les courbes relient les tailles d'échantillon optimales pour chaque exposant q ; elles sont tracées sur l'échelle linéaire (à gauche) et sur l'échelle logarithmique (à droite). Les tailles de population sont inscrites sur la barre horizontale au bas de chaque graphique. Sur l'échelle logarithmique, les courbes sont linéaires. Cette échelle produit aussi une répartition plus uniforme des tailles de population des cantons.

Il en est ainsi pour les plans d'échantillonnage stratifiés dans lesquels les strates coïncident avec les petits domaines. À la section 4, nous discutons des plans d'échantillonnage pour lesquels ce genre de contrôle ne peut être exercé; ces plans pour tous les sous-domaines, mais aussi quand les paramètres visés par l'inférence sont les moyennes de petit domaine. Si l'on veut estimer des totaux de population, l'équitépartition de l'échantillon entre les sous-domaines n'est pas efficace, parce qu'elle pénalise l'estimation pour les petits domaines les plus peuplés. Même si l'on estime les proportions ou des taux (pourcentages), les variances intradomaine dépendent de la proportion de population, quoique la dépendance soit faible lorsque toutes les proportions sont loin de zéro et de l'unité. Pour des travaux plus récents sur les plans d'échantillonnage pour l'estimation pour petits domaines, voir Marker (2001).

2. Plan optimal pour l'estimation directe

Nous résolvons le conflit entre les objectifs d'estimation efficace de paramètres au niveau du petit domaine θ_d en choisissant le plan d'échantillonnage à ce niveau qui minimise la somme pondérée des variances d'échantillonnage (EQM),

$$(1) \quad \min_n \sum_{d=1}^D P_d v_d,$$

sachant que la taille globale d'échantillon $n = \sum_{d=1}^D n_d$ est fixe; $\mathbf{1}_D$ est le vecteur des unités de longueur D . Le coefficient P_d est nommé *priorité d'inférence*. Une valeur plus grande de P_d (par rapport aux valeurs $P_{d'}, d' \neq d$) implique qu'il est plus important de réduire v_d , parce que l'augmentation de la contribution du petit domaine d à la somme (1) est plus importante que pour les autres petits domaines. Le problème d'optimisation (1) est résolu par la méthode des multiplicateurs de Lagrange, ou simplement par substitution de $n_1 = n - n_2 - \dots - n_D$, si bien qu'il comporte alors $D - 1$ variables fonctionnellement non corrélées. La solution satisfait la condition

$$P_d \frac{\partial v_d}{\partial n_d} = \text{const.}$$

En général, il n'est pas possible d'obtenir une expression analytique des tailles optimales des sous-échantillons n_d , mais si $v_d = \sigma_d^2/n_d$, comme dans le cas de l'échantillonnage aléatoire simple à l'intérieur des petits domaines, la solution est proportionnelle à $\sigma_d \sqrt{P_d}$, c'est-à-dire

$$n_d^* = n \frac{\sigma_d \sqrt{P_d}}{\sigma_1 \sqrt{P_1} + \dots + \sigma_D \sqrt{P_D}}.$$

Lorsque les variances intra domaine σ_d^2 sont égales, $\sigma_1^2 = \dots = \sigma_D^2 = \sigma^2$, la solution se simplifie encore davantage; les tailles optimales d'échantillon sont proportionnelles à $\sqrt{P_d}$ et ne dépendent pas de σ^2 .

Dans la plupart des contextes, il est difficile d'exprimer un ensemble approprié de priorités P_d et il est donc plus constructif de proposer une classe paramétrique commune de priorités $\mathbf{P} = (P_1, \dots, P_D)^T$ et d'illustrer son effet sur la répartition de la taille de l'échantillon. Nous proposons les priorités $P_d = N_d^q$ pour $0 \leq q \leq 2$. Si $q = 0$, l'inférence est de même importance pour chaque petit domaine. À mesure

sous-échantillons de même taille sont attribués à chaque petit domaine, lorsque les variances dans les sous-domaines sont égales, que la correction pour population finie peut être ignorée et que les coûts d'enquête par sujet sont les mêmes pour tous les sous-domaines, mais aussi quand les paramètres visés par l'inférence sont les moyennes de petit domaine. Si l'on veut estimer des totaux de population, l'équitépartition de l'échantillon entre les sous-domaines n'est pas efficace, parce qu'elle pénalise l'estimation pour les petits domaines les plus peuplés. Même si l'on estime les proportions ou des taux (pourcentages), les variances intradomaine dépendent de la proportion de population, quoique la dépendance soit faible lorsque toutes les proportions sont loin de zéro et de l'unité. Pour des travaux plus récents sur les plans d'échantillonnage pour l'estimation pour petits domaines, voir Marker (2001).

La présente section se termine par une description de la notation utilisée dans la suite de l'article. Nous supposons que les paramètres de population au niveau du petit domaine $\theta_d, d = 1, \dots, D$, sont estimés par $\hat{\theta}_d$ avec des erreurs quadratiques moyennes (EQM) v_d respectives qui sont des fonctions des tailles des sous-échantillons dans les petits domaines $n_d, v_d = v(n_d)$. La taille globale de l'échantillon est dénotée par n et nous supposons qu'elle est fixe. Les tailles de population sont dénotées par N (globale) et N_d (pour le petit domaine d). Par souci de concision, nous dénotons $\mathbf{n} = (n_1, \dots, n_D)^T$. La plupart des paramètres de population θ sont des fonctions d'une seule variable, le total et ainsi de suite. La variable peut être enregistrée directement durant le sondage ou construite d'après une ou plusieurs variables directes. Bien que notre développement ne soit pas limité à ce genre de paramètres, la justification est plus simple en ce qui les concerne. Nous disons qu'un estimateur de θ_d est *direct* s'il s'agit d'une fonction de la variable étudiée sur les sujets du petit domaine d seulement.

Nous supposons que chaque estimateur direct envisagé est sans biais. Cette hypothèse n'est pas particulièrement restrictive, car la plupart des estimateurs directs sont des estimateurs naïfs ou étroitement reliés à ces derniers. Nous supposons que les tailles d'échantillon pour les petits domaines sont sous le contrôle du concepteur de l'enquête.

Calcul de la taille de l'échantillon pour l'estimation pour petits domaines

Nicholas Tibor Longford

Résumé

Nous décrivons une approche générale de détermination du plan d'échantillonnage des enquêtes planifiées en vue de faire des inférences pour de petits domaines (sous-domaines). Cette approche nécessite la spécification des priorités d'inférence pour les petits domaines. Nous établissons d'abord des scénarios de répartition de la taille de l'échantillon pour l'estimateur direct, puis pour les estimateurs composites et bayésien empirique. Nous illustrons les méthodes à l'aide d'un exemple de planification d'un sondage de la population suisse et d'estimation de la moyenne ou de la proportion d'une variable pour chacun des 26 cantons.

Mots clés : Efficacité; estimation pour petits domaines; priorité d'inférence; répartition de la taille de l'échantillon.

1. Introduction

Le plan d'échantillonnage est un instrument essentiel à la production d'estimations efficaces et d'autres formes d'inférence au sujet d'une grande population, lorsque les ressources disponibles ne permettent pas de recueillir l'information pertinente pour chaque membre de la population. Dans ce contexte, nous interprétons l'efficacité comme étant la combinaison optimale d'un plan d'échantillonnage et d'un estimateur d'un paramètre de population θ . Par optimalité, nous entendons que l'erreur quadratique moyenne est minimale, quoique le développement présenté dans l'article puisse être adapté à d'autres critères. Le groupe de plans de sondage possibles est défini par les ressources et celles-ci sont habituellement exprimées en fonction d'une taille fixe d'échantillon. Cette approche n'est pas toujours appropriée, parce que les coûts moyens par sujet ne sont pas nécessairement les mêmes pour tous les plans d'échantillonnage. Toutefois, si nous considérons une gamme limitée de plans, nous pouvons ignorer ce point.

Le problème de l'établissement du plan d'échantillon-nage afin d'estimer efficacement une grandeur unique est bien compris et des solutions existent pour bon nombre de spécifications utilisées fréquemment. La plupart comportent un problème d'optimisation univarié sous contraintes. L'établissement du plan d'échantillonnage pour l'estimation de plusieurs paramètres est considérablement plus complexe, parce que le problème comprend plusieurs facteurs, habituellement un pour chaque paramètre. Il est essentiel d'optimiser le plan simultanément pour tous les facteurs, parce que les objectifs d'inférence efficace au sujet des paramètres peuvent être conflictuels. Par exemple, dans l'estimation pour petits domaines, l'allocation d'une part plus généreuse de la taille de l'échantillon à un petit domaine doit être compensée par une allocation moins généreuse à un ou à plusieurs autres.

Au cours des dernières décennies, la production de statistiques pour des petits domaines est devenue un important sujet de recherche en méthodologie d'enquête (Fay et Herriot 1979; Platek, Rao, Särndal et Singh 1987; Ghosh et Rao 1994; Longford 1999; Rao 2003), étant donné l'intérêt grandissant des organismes gouvernementaux, du secteur de la publicité et du marketing et de celui de la finance et des assurances pour ce genre d'information. À l'heure actuelle, de nombreuses enquêtes à grande échelle sont conçues en vue de produire des estimations de niveau national, mais sont parfois utilisées après coup pour faire des inférences au sujet de petits domaines. Cela n'aurait pas d'inconvénient si les plans d'échantillonnage optimaux pour l'inférence sur petits domaines et l'inférence nationale étaient les mêmes. Nous montrons dans le présent article qu'il n'en est pas ainsi et que le plan d'échantillonnage peut effectivement être cible pour l'estimation pour petits domaines, en tenant compte de l'objectif de production d'estimations efficaces de paramètres de niveau national. Pour éviter le cas banal, supposons que les populations des petits domaines soient de taille inégale. Nous appliquons les méthodes au problème de la planification d'inférences au sujet des 26 cantons de la Suisse; la taille de la population de ces cantons varie de 15 000 (Appenzell-Innerrhodén) à 1,23 million (Zürich). La population de la Suisse se chiffre à 7,26 millions d'habitants.

La littérature traitant de la planification des enquêtes pour l'estimation pour petits domaines est peu abondante. L'une des contributions importantes est celle de Singh, Gambino et Mantel (1994). Dans l'une des approches dont discutent ces auteurs, la taille prévue de l'échantillon de l'Enquête sur la population active du Canada est divisée en deux. Une partie est répartie optimalement en vue de la production d'estimations de niveau national (domaine) et l'autre est répartie optimalement en vue de l'estimation pour petits domaines (sous-domaines). Pour ce dernier objectif, des

Les valeurs lissées pour chaque strate sont les éléments du vecteur Xw . Un problème similaire peut être posé pour les non-répondants, lorsqu'on veut lisser les effets du plan de sondage.

7. Conclusion

Il y a beaucoup de chevauchement entre les deux objectifs traditionnels de la répartition de l'échantillon de la CVD, qui sont d'obtenir une variance minimale pour le taux national estimé de sous-dénombrement (objectif I) et d'obtenir des variances égales pour les taux de sous-dénombrement estimés de chaque province (objectif II), et les deux objectifs additionnels examinés dans cet article, qui sont de réduire au minimum la variance du paiement de péréquation total (objectif III) et de produire des CV égaux pour l'estimation de la population de chaque province bénéficiaire (objectif IV). Néanmoins, la prise en compte explicite de ces deux objectifs additionnels peut permettre à la taille de l'échantillon pour le Québec et pour l'Alberta de varier indépendamment de ceux des autres provinces. La méthode proposée dans cet article pour réaliser un compromis entre différentes répartitions, optimal en ce qui a trait aux différents objectifs, consiste à prendre, pour chaque province, la taille maximale de l'échantillon sur chacune des répartitions. Cette méthode fournit une justification plus directe de la répartition.

Une comparaison des erreurs types du SCE et des erreurs-types résultant de la formule approximative (3.6) montre que, pour la CVD de 2001, n unités échantillonnées

Bibliographie

Brackstone, G.J., et Rao, J.N.K. (1976). Raking ratio estimators. *Survey Methodology*, 2, 63-69.

Clark, C. (septembre 2000). 2001 Reverse Record Check: Provincial and Territorial sample allocation. Document non publié. Ottawa. Statistique Canada.

Deming, W.E., et Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11, 427, 444.

Deville, J.-C., et Samdal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

Grce une mthode de rpartition qui utilise (5.1) et un tableau comme le tableau 5.2, on peut voir clairement la raison d'tre de la taille de l'chantillon d'une province. Par exemple, en examinant la rpartition finale de l'chantillon au tableau 5.2, si on estime que 5 867 observations au Manitoba ne sont pas suffisantes, alors il faut prciser l'objectif pour lequel elles ne sont pas suffisantes. S'il s'agit d'apporter une amlioration pour l'objectif II (ou l'objectif IV), alors il faut ggalement accrotre la taille de l'chantillon dans toutes les provinces de l'Atlantique et dans toutes les provinces de l'Ouest (ou dans toutes les provinces de l'Atlantique et dans toutes les provinces de l'Ouest sauf l'Alberta).

6. Rpartition intraprovinciale de l'chantillon

Bien que l'quation (3.5) permette de voir que la rpartition intraprovinciale de l'chantillon dans une province de slection affecte la variance des estimations pour les autres provinces, nous n'essayerons d'optimiser la rpartition dans une province que pour l'estimation dans cette mme province. Autrement dit, notre problme pour chaque province p , consiste a minimiser

$$(6.1) \sum_{h \in \left\{ \begin{smallmatrix} \text{strates de la province} \\ \text{de slection } p \end{smallmatrix} \right\}} \frac{D_h^h U^{hp} (N^h - U^{hp})}{n_h}$$

sous la contrainte

$$\sum_{h \in \left\{ \begin{smallmatrix} \text{strates de la province} \\ \text{de slection } p \end{smallmatrix} \right\}} n_h = n^p$$

ou n^p est une taille totale d'chantillon dtermine antieurement pour la province p . Notons que la taille d'chantillon alloue a la base des personnes omises est fixe, si bien que nous ne tiendrons pas compte dans la suite, des strates de cette base de sondage et que n^p n'inclut pas la taille de l'chantillon tir de la base des personnes omises. La solution de ce problme de minimisation est

$$(6.2) n_h^h = n^p \frac{\sqrt{D_h^h U^{hp} (N^h - U^{hp})}}{\sum_{h \in \left\{ \begin{smallmatrix} \text{strates de la province} \\ \text{de slection } p \end{smallmatrix} \right\}} \sqrt{D_h^h U^{hp} (N^h - U^{hp})}}$$

pour chaque strate h^* de la province de slection p .

Il a t vu a la section 4 que des donnes empiriques au niveau provincial montrent que le facteur D_h^h est inversement proportionnel au cube du taux de rponse a la CVD. Pour la rpartition de l'chantillon de 2001, on a suppos que D_h^h varierait comme l'inverse du taux de rponse. Pour limiter le dplacement de l'chantillon comparativement a celui de 2001, c'est--dire passer de strates a taux de rponse lev, comme celles de la base du recensement et de la base des naissances, a des strates a taux de rponse

faible, comme celles de la base des immigrants ou de la base des rsidents non permanents, nous rendons D_h^h inversement proportionnel au carr de du taux de rponse dans la strate h . Notons qu'a la diffrence de ce qu'on a du supposer a la section 3, ici le facteur D_h^h sert a compenser la non-rponse seulement, il n'a pas a compenser pour la stratification puisqu'il est dfini au niveau de la strate. Ceci milite aussi en faveur d'un facteur infrieur a l'inverse du cube du taux de rponse.

Comme pour la rpartition de l'chantillon de 2001, nous devons rsoudre le problme de la projection fiable des valeurs de 2006 de U^{hp} et D_h^h pour chaque strate h . Puisque les bases des naissances, des immigrants et des rsidents non permanents ne contiennent chacune qu'une seule strate par province, nous proposons d'utiliser pour ces strates leurs tailles de 2006, ainsi que les taux de rponse et de personnes omises de 2001, en apportant, si ncessaire, des corrections spciales pour les provinces les moins peuples. Une procdure comparable peut tre utilise pour les strates des rserves indiennes de la base du recensement. Les autres strates de la base du recensement sont formes a partir du sexe, de l'tat matrimonial (marie(e), non marie(e)), et du groupe d'ge des units. Pour ces strates, en utilisant les mmes groupes d'ge selon le sexe et l'tat matrimonial, il serait possible, pour chaque province, de calculer par ajustement proportionnel itratif (raking), les projections *nationales* sur des valeurs de marge donnes par des projections provinciales, et d'utiliser les valeurs ainsi calcues dans l'quation (6.2). Plus prcisement, pour calculer les projections de U^{hp} pour toutes les strates h de la province de slection p , nous commenons par calculer, en nous fondant sur les taux estims de 2001 et les tailles des strates de 2006, une projection du nombre de personnes omises, classes dans la province o elles ont t slectionnes, pour chaque cellule (sexe \times tat matrimonial \times groupe d'ge). Ces chiffres *nationaux* couvrent les cellules d'une matrice tridimensionnelle. Toujours en nous fondant sur les taux estims de 2001 et les tailles des strates de 2006, nous pourrions obtenir les projections pour U^{hp} dont la somme correspond aux totaux provinciaux souhaits selon le sexe, selon l'tat matrimonial et selon le groupe d'ge. Afin d'viter les problmes de convergence, de simplifier la programmation et de rendre le processus plus flexible, l'ajustement proportionnel itratif (raking) sera remplac par la rsolution d'un problme de calage. En fait, la plus

$$n_p = \max(n_{pI}, n_{pII}, n_{pIII}, n_{pIV}) \quad p = 1, \dots, 10. \quad (5.1)$$

Qu'on utilise le maximum des 4 tailles comme dans (5.1), une moyenne arithmétique pondérée, ou une moyenne géométrique pondérée, chaque méthode fait usage de 4 paramètres arbitraires (3 si la taille totale d'échantillon est fixe). Pour la méthode du maximum, des valeurs relatives plus élevées de n_I (respectivement n_{II} , n_{III} et n_{IV}) donnent une plus grande importance relative à l'objectif I (respectivement II, III et IV).

Le tableau 5.2 donne un exemple avec $n_I = 30\,000$, $n_{II} = 64\,000$, $n_{III} = 25\,000$ et $n_{IV} = 48\,078$. La taille totale de l'échantillon qui résulte est 70 028. Les chiffres en caractères gras sont égaux au maximum sur les quatre répartitions, n_p . De petits changements apportés à n_{III} n'auraient un effet que sur la répartition finale pour le Québec. Cela indique qu'avec les tailles d'échantillon n_I , n_{II} , n_{III} et n_{IV} choisies ci-dessus, la taille finale de l'échantillon attribué au Québec est dictée par l'objectif III : une estimation précise du paiement de pétrogation totale. De même, la taille finale de l'échantillon attribué à l'Ontario est dictée par l'objectif I : une estimation précise du taux

national de sous-dénombrement. La taille finale de l'échantillon attribué à l'Alberta est dictée par l'objectif II : variances égales pour le taux estimé de sous-dénombrement de chaque province. Les tailles finales des échantillons des autres provinces sont aussi bien dictées par l'objectif II que par l'objectif IV (précision égale du paiement de pétrogation estimé par personne). Comme nous l'avons signalé à la section 3, n_{pII}/n_{pIV} est constant pour les huit provinces bénéficiaires. Dans l'exemple ci-dessus, à cause du choix «judicieux» de n_{IV} , la valeur de la constante est un. Une diminution apportée à n_{IV} entraînerait une diminution de la taille finale de l'échantillon de l'Alberta, mais non de ceux des autres provinces. Nous observons également que n_{pI}/n_{pIII} ne varie pas beaucoup pour les huit provinces bénéficiaires. L'ajout des objectifs III et IV (se rapportant aux paiements de pétrogation) permet de contrôler séparément la taille de l'échantillon du Québec et de celui de l'Alberta. Lorsque seuls les objectifs I et II étaient utilisés, la taille de l'échantillon du Québec avait tendance à être étroitement liée à celle de l'Ontario tandis que la taille de l'échantillon de l'Alberta était étroitement liée à celle des échantillons des autres provinces.

Tableau 5.1
Valeurs des paramètres

Province	N_p	D_p	P_p	R_p	$R_{p^{ben}}$
T.-N.-L.	551 987	1,0804	524 722	0,0339	0,0464
I.-P.-E.	145 173	1,0882	132 473	0,0334	0,0307
N.-É.	995 651	1,1527	947 099	0,0492	0,0464
N.-B.	797 488	1,1345	736 129	0,0493	0,0466
Qc	8 079 167	1,1740	7 381 352	0,0510	0,0471
Ont.	13 423 132	1,2752	11 702 797	0,0653	0,0565
Man.	1 262 547	1,1558	1 136 146	0,0466	0,0437
Sask.	1 082 238	1,1223	996 562	0,0437	0,0430
Alb.	3 373 128	1,2478	3 010 105	0,0490	0,0403
C.-B.	4 570 444	1,3369	4 014 502	0,0761	0,0669
Can.	34 280 955	1,2039	30 581 887	0,0587	0,0524

Tableau 5.2
Répartition de l'échantillon au niveau des provinces avec
 $n_I = 30\,000$, $n_{II} = 64\,000$, $n_{III} = 25\,000$, et $n_{IV} = 48\,078$

Province	n_{pI}	n_{pII}	n_{pIII}	n_{pIV}	n_p	n_{pI}/n_{pIII}
T.-N.-L.	427	3 816	546	3 816	3 816	0,78
I.-P.-E.	96	3 956	132	3 956	3 956	0,73
N.-É.	796	5 822	1 107	5 822	5 822	0,72
N.-B.	634	5 921	881	5 921	5 921	0,72
Qc	6 562	6 399	9 262	6 399	9 262	0,71
Ont.	12 385	9 220	3 148	12 385	12 385	3,93
Man.	982	5 867	1 331	5 867	5 867	0,74
Sask.	823	5 234	1 139	5 234	5 234	0,72
Alb.	2 622	6 702	1 015	0	6 702	2,58
C.-B.	4 673	11 063	6 440	11 063	11 063	0,73
Total	30 000	64 000	25 000	48 078	70 028	

que l'effet de plan de sondage des équations (3.5) et (3.6) est approximativement égal à l'inverse du cube du taux de réponse. Ceci semble indiquer qu'un échantillon de « n » unités avec un taux de réponse « r » correspond à un échantillon de n × r³ unités plutôt qu'à la taille attendue de n × r unités, à cause de la concentration des non-répondants parmi les personnes omises par le recensement. Le SCE tient compte que les personnes omises sont moins susceptibles de répondre. Cette érosion de la précision due à la non-réponse se produit même si le plan d'échantillonnage est stratifié de façon plus efficace que le plan d'échantillonnage d'une strate par province que nous avons supposé.

Tableau 4.1 Comparaison de l'erreur-type

Province	Taux de réponse	D = (taux de réponse) ³	Erreur-type	Erreur-type
E.-T.	de l'esti-	mation des	omission des	omission du
(3.6)	/			
E.-T.	Erreur-type			GES
1,06	1 689	1 783	1 689	1,06
1,39	734	1 021	3 903	0,99
0,99	3 955	3 903	3 229	1,01
1,01	3 229	3 272	19 664	1,01
1,00	31 602	31 502	19 915	1,00
0,93	5 115	4 762	5 115	0,93
1,02	3 840	3 921	3 840	1,02
1,00	10 505	10 493	14 619	0,99
0,99	14 763	14 619	42 041	1,00
Can.	0,94	1,20	42 074	

Aucune étude semblable n'a été faite pour comparer l'effet de plan et le taux de non-réponse lors des CVD précédentes. La méthode du rajustement des poids pour compenser la non-réponse est différente, et la nature même de la non-réponse est significativement différente de ce qu'elle était avant 2001.

5. Répartition finale de l'échantillon au niveau des provinces et exemple

Le tableau 5.1 montre les valeurs des paramètres qui seront utilisées dans l'exemple. Les valeurs de N_p sont des projections de la taille de la base de sondage de la CVD pour 2006; les autres paramètres sont fondés sur les données de la CVD de 2001.

Comme nous pourrions nous y attendre, les valeurs de R_p en Alberta et en Ontario montrent que seulement un petit nombre des unités sélectionnées dans ces deux provinces sont classées comme omises par le recensement dans les provinces bénéficiaires.

La taille de l'échantillon final attribué à la province p est simplement

$$n^{pIV} = m^{IV} \left[\frac{D^p R^p (N^p / P^p - R^p)}{\sum_{d=1}^p D^d R^d (N^d / P^d - R^d)} \right] \quad p = 1, \dots, 8 \quad (3.13)$$

L'objectif IV de l'échantillon de taille totale n^{IV} est : $(P^p - R^p)$. Par conséquent, la répartition optimale pour est proportionnel à $(1/P^p) D^p U^p (N^p - U^p) = D^p R^p (N^p / P^p - R^p)$. Les provinces bénéficiaires seront alors égales si $n_h = p$, pour $h = p$, coefficients de variation des estimations de la population des $U^{hp} = 0$ pour $p \neq h$ et que $U^{dp} = U^p$. Les coefficients de variation des estimations de la population des $U^{hp} = 0$ pour $p \neq h$, spécifiquement si p est une petite province. Comme nous l'avons fait pour l'objectif II, nous supposons plutôt que U^{hp} pour obtenir des estimations suffisamment précises des U^{hp} pour $p \neq h$, ici aussi, une deuxième difficulté consiste à province bénéficiaire. Ce problème a huit équations à huit

avec les deux provinces non bénéficiaires ayant $n^{pIV} = 0, p = 9, 10$.

Il convient de signaler que n^{pIV} est constant pour les huit provinces bénéficiaires. Cela montre un important chevauchement de l'objectif II (précision égale des taux provinciaux estimés de sous-dénombrement), qui est un objectif traditionnel de la répartition de l'échantillon de la CVD, et de l'objectif IV (précision égale des paiements de péréquation par personne des provinces bénéficiaires). Comme nous le constaterons à la section 5, n^{pI}/n^{pIII} , pour les huit provinces bénéficiaires, est pratiquement constant et également. Il en ressort un important chevauchement de l'objectif I (précision maximale du taux national estimé de sous-dénombrement), un objectif traditionnel de la répartition de l'échantillon de la CVD, et de l'objectif III (précision maximale du total des paiements de péréquation).

4. Effet de plan

L'erreur-type des estimations de la CVD de 2001 a été calculée par le système généralisé d'estimation (SGE). Cette erreur-type prend en compte le plan de sondage de la CVD et la non-réponse à l'enquête en supposant que les répondants sont sélectionnés au moyen d'un plan de sondage à plusieurs degrés. Le tableau 4.1 présente une comparaison de l'erreur-type obtenue de (3.6) avec celle calculée par le SGE. Pour cette comparaison, un effet de plan de sondage égal à l'inverse du cube du taux de réponse de la province de sélection a été utilisé.

Nous voyons que l'erreur-type pour l'Ile-du-Prince-Édouard dérivée de (3.6) est supérieure de 39 % à celle calculée par le SGE; ceci est attribuable à une observation aberrante qui a un effet plus important sur l'estimation à partir de (3.6) que sur l'estimation du SGE. Pour la plupart des provinces, l'erreur-type (3.6) est proche de celle calculée par le SGE. Ces résultats empiriques montrent donc

où $P_{\cdot} = \sum_{d=1}^{10} P^d$, $R_{\cdot} = U / P_{\cdot}$, $U_{\cdot} = U / P$ et $U_h = \sum_{d=1}^{10} U_{h,d}$. Cette variance du taux national estimé de sous-dénombrement sera réduite au minimum si n_h est proportionnel à $\sqrt{D_h U_h (N_h - U_h)} = N_h \sqrt{D_h R_h (1 - R_h)}$, où $R_h = U_h / N_h$. Par conséquent, la répartition optimale pour l'objectif I d'un échantillon de taille totale n_I est :

$$n_{I1} = n_I \left[\frac{N^p \sqrt{D^p R^p (1 - R^p)}}{\sum_{d=1}^p N^d \sqrt{D^d R^d (1 - R^d)}} \right] \quad p = 1, \dots, 10. \quad (3.8)$$

Cette formule représente une améloration par rapport à celle utilisée pour la CVD de 2001 (voir Clark (2000), où aucun effet du plan de sondage n'était appliqué à la partie de l'échantillon répartie de manière à produire la meilleure estimation au niveau du Canada. En outre, pour la CVD de 2001, n_p était proportionnel à la population prévue dans la province p . Il est correct que n_p dépende plutôt de la taille des bases de sondage provinciales; il est aussi correct qu'il dépende de la répartition provinciale du sous-dénombrement.

Objectif II :

Nous pouvons utiliser l'équation (3.5) pour déterminer quelles valeurs de n_h donnent la même variance pour les taux provinciaux estimés de sous-dénombrement. Ce problème a dix équations à dix inconnues. Une deuxième difficulté consiste à obtenir des estimations suffisamment précises des U_{hp} pour $p \neq h$, spécialement si p est une petite province. Bien qu'il soit souvent raisonnable de croire que le taux de personnes omises dans une petite province p , $R^p = U^p / P^p$, qui est observé lors d'un recensement soit une bonne prédiction du taux qui sera observé au prochain recensement, les valeurs individuelles des U_{hp} pour $p \neq h$ sont plus difficiles à estimer et encore plus à prédire. Nous supposons plutôt que $U_{hp} = 0$ pour $p \neq h$ et que $U^{pp} = U^p$, ce qui aura pour effet de mitiger l'effet de données aberrantes sur les variances attendues. Les estimations provinciales du taux de sous-dénombrement seront alors de variance égale, si n_h , pour $h = p$, est proportionnel à $(1/P^p) D^p U^p (N^p - U^p) = D^p R^p (N^p / P^p - R^p)$. Par conséquent, la répartition optimale pour l'objectif II de l'échantillon de taille totale n_{II} est :

$$n_{II1} = n_{II} \left[\frac{D^p R^p (N^p / P^p - R^p)}{\sum_{d=1}^p D^d R^d (N^d / P^d - R^d)} \right] \quad p = 1, \dots, 10. \quad (3.9)$$

Il convient de souligner que pour la CVD de 2001, pour la partie de l'échantillon répartie de manière à assurer

l'égalité sur le plan de la précision des estimateurs provinciaux, les tailles des échantillons ont été fixées proportionnellement à $D^p R^p (1 - R^p)$ (voir Clark (2000)). Utiliser N^p / P^p au lieu de 1, permet de tenir compte des unités figurant dans la base de sondage de la province qui quittent la population de la province, ainsi que des unités de la population de la province qui ne font pas partie de la base de sondage pour cette province, de sorte que l'effet de plan d'échantillonnage. En 2001, une correction pour les unités sortant de la population était apportée par le biais de l'effet de plan, et aucune correction n'a été apportée pour les unités de la population ne faisant pas partie de la base.

Objectif III :

L'estimation de la population totale des provinces bénéficiaires a une variance égale à

$$V(\hat{P}_{\cdot}^{\text{ben}}) = V(\hat{U}_{\cdot}^{\text{ben}}) \approx \sum_{h=1}^H D^h U^h_{\text{ben}} (N^h - U^h_{\text{ben}}) n_h \quad (3.10)$$

où $P_{\cdot}^{\text{ben}} = \sum_{d=1}^8 P^d$, $U_{\cdot}^{\text{ben}} = \sum_{d=1}^8 U^d_{hp}$ et $U^{\text{ben}}_{hp} = \sum_{d=1}^8 U^d_{hp}$ sont des sommes sur les huit provinces bénéficiaires (on suppose que les provinces bénéficiaires sont numérotées avec $p = 1, \dots, 8$, et que les provinces non bénéficiaires sont numérotées avec $p = 9, 10$). L'équation (3.10) est réduite au minimum si n_h , pour $h = 1, \dots, 10$, est proportionnel à $\sqrt{D^h U^h_{\text{ben}} (N^h - U^h_{\text{ben}})} = N^h \sqrt{D^h R^h_{\text{ben}} (1 - R^h_{\text{ben}})}$ où $R^h_{\text{ben}} = U^h_{\text{ben}} / N^h$. Par conséquent, la répartition optimale pour l'objectif III d'un échantillon de taille totale n_{III} est :

$$n_{III1} = n_{III} \left[\frac{N^p \sqrt{D^p R^p_{\text{ben}} (1 - R^p_{\text{ben}})}}{\sum_{d=1}^p N^d \sqrt{D^d R^d_{\text{ben}} (1 - R^d_{\text{ben}})}} \right]$$

$$p = 1, \dots, 10. \quad (3.11)$$

Signalons qu'étant donné que les unités sélectionnées dans une province peuvent être classées dans une autre province, R^p_{ben} et n_{III} ne sont pas nécessairement nuls lorsque p est une province non bénéficiaire.

Objectif IV :

À partir de l'équation (3.6), nous obtenons :

$$CV(\hat{P}_{\cdot}^p) \approx \frac{1}{D^p} \sqrt{\sum_{h=1}^{10} D^h U^h_{hp} (N^h - U^h_{hp}) n_h} \quad (3.12)$$

Nous pouvons utiliser cette formule pour déterminer quelles valeurs de n_h donnent le même coefficient de variation pour les estimations de la population de chaque

Tableau 2.1 Les quatre objectifs de la répartition provinciale de l'échantillon

Objectif	Description (description équivalente)
I	Réduire au minimum la variance pour le taux national estimé de sous-dénombrement. (Réduire au minimum le CV de l'estimation nationale de la population.)
II	Produire des variances égales pour les taux estimés de sous-dénombrement dans chaque province. (Produire des estimations démographiques provinciales dont le CV est égal.)
III	Réduire au minimum la variance du paiement de péréquation totale. (Réduire au minimum la variance de l'estimation de la population totale des provinces bénéficiaires.)
IV	Produire des variances égales pour le paiement de péréquation par personne dans chaque province bénéficiaire. (Produire des CV égaux pour l'estimation de la population de chaque province bénéficiaire, ou encore, produire des variances égales pour les taux estimés de sous-dénombrement des provinces bénéficiaires.)

Dans la présente section, nous fournissons d'abord quelques précisions sur la notation adoptée, puis nous présentons des formules de variance approximatives pour les estimations démographiques et les taux de sous-dénombrement estimés. Nous examinons la question de l'optimalité relativement aux quatre objectifs susmentionnés.

Cinq bases de sondage sont utilisées aux fins de la CVD dans les provinces : la base du recensement (personnes dénombrees au recensement précédent), la base des naissances (naissances intercensitaires), la base des immigrants (immigrants intercensitaires), la base des résidents non permanents et la « base des personnes omises ». La « base des personnes omises » se compose des personnes échantillonnées de la CVD précédente qui ont été omises au recensement précédent. Avec leurs poids, ils représentent la sous population de personnes dénombreables qui ne sont pas couvertes par l'une des quatre autres bases. Chaque base dans chaque province est stratifiée séparément. Un échantillon aléatoire stratifié est sélectionné dans chaque base. Toutes les personnes de la base des personnes omises sont incluses dans l'échantillon.

Soit U_{hp} le nombre de personnes omises dans la strate h qui sont classées dans la province (de classification) p . De la même façon, soient E_{hp} et O_{hp} , respectivement, le nombre de personnes dénombrees et le nombre de personnes surdénombrees dans la strate h qui sont classées dans la province p , et $P_{hp} = U_{hp} + E_{hp} - O_{hp}$. Le taux de sous-dénombrement pour la province p peut alors s'écrire :

3. Répartition optimale de l'échantillon au niveau provincial

où $U^p = \sum^h U_{hp}$ et $P^p = \sum^h P_{hp}$. Nous voyons que P^p est égal à P^p tel que défini à la section précédente.

Un estimateur du taux de sous-dénombrement pour la province p est

$$\hat{R}^p = U^p / P^p, \tag{3.2}$$

où U^p et P^p sont des estimateurs de U^p et P^p , respectivement. Une linéarisation donne

$$V(\hat{R}^p) \approx \frac{1}{U^2} \left[P^p V(U^p) + \frac{D^p}{U^2} V(P^p) \right]. \tag{3.3}$$

Le deuxième terme entre crochets étant négligeable par rapport au premier, nous avons

$$V(\hat{R}^p) \approx \frac{1}{U_{hp}^2} \sum^h \frac{D^p}{U_{hp}^2} \frac{n_h}{(N_h - U_{hp})}, \tag{3.4}$$

où N_h est la taille de la strate h , et n_h est la taille de l'échantillon dans la strate h . Cette expression ne tient pas compte du facteur de correction pour population finie. Pour ce qui suit, nous supposons qu'il n'y a pas de non-réponse et nous supposons qu'il y a une seule strate par province de sélection (pas de stratification selon la base, l'âge, le sexe, etc.). Cette dernière supposition sera bien sûr abandonnée à la section 6 qui traite de la répartition de l'échantillon aux strates infraprovinciales. Pour compenser les effets de la stratification infraprovinciale et de la non-réponse, nous introduisons un effet de plan, D_h . Nous supposons que cet effet de plan ne varie qu'avec la strate h , en particulier, le même effet de plan est utilisé pour exprimer la variance de l'estimateur du nombre de personnes sélectionnées dans la strate h qui sont omises dans la province p , quel que soit p . Une approximation de la variance (3.4) peut être donnée par

$$V(\hat{R}^p) \approx \frac{1}{U_{hp}^2} \sum^h \frac{D_h^p}{U_{hp}^2} \frac{n_h}{(N_h - U_{hp})}, \tag{3.5}$$

et

$$V(U^p) \approx \sum^h \frac{D_h^p}{U_{hp}^2} \frac{n_h}{(N_h - U_{hp})}, \tag{3.6}$$

où, cette fois, les sommations sont faites sur les provinces de sélection.

Objectif 1:

À partir de l'équation (3.5), nous obtenons :

$$V(\hat{R}^p) \approx \frac{1}{U_{hp}^2} \sum_{h=1}^h \frac{D_h^p}{U_{hp}^2} \frac{n_h}{(N_h - U_{hp})}, \tag{3.7}$$

Nous constatons que la population de l'Alberta n'a pas d'incidence sur le paiement de péréquation d'une province bénéficiaire. La population de l'Ontario a un effet sur le paiement de péréquation seulement par F^{norme} . Dans le cas des provinces de l'Atlantique, leur paiement de péréquation varie linéairement en fonction de leur population, puisque celle-ci n'a pas d'effet sur F^{norme} . Si nous supposons que F^{norme} est connu, alors nous pouvons dire que, pour toute province bénéficiaire, une erreur d'une personne dans sa population a un effet de $C^{\text{norme}}/P^{\text{norme}}$ dollars sur son paiement de péréquation. Cela ne veut pas dire que le paiement de péréquation d'une province bénéficiaire dépend seulement de sa population et non de la population des provinces de référence. Toutefois, comme nous pourrions le constater, la plus grande partie de l'erreur d'échantillonnage dans le paiement de péréquation est attribuable à l'erreur d'échantillonnage dans l'estimation de la population de la province bénéficiaire et une partie relativement petite, à l'erreur d'échantillonnage dans l'estimation de la population des provinces de référence.

Si les symboles avec un chapeau représentent des estimateurs, alors d'après (2.2),

$$V(\hat{E}^p) \approx C^2 \frac{1}{P^{\text{norme}}{}^2} + \left(\frac{P^{\text{norme}}}{P} \right)^2 V(\hat{P}^{\text{norme}}) - 2 \frac{P^{\text{norme}}}{P} \text{Cov}(\hat{P}^p, \hat{P}^{\text{norme}}) \quad (2.3)$$

Comme nous stratifions séparément pour chaque province, pour une province bénéficiaire p , qui n'est pas l'une des provinces de référence, nous avons, en faisant abstraction de la migration interprovinciale, $\text{Cov}(\hat{P}^p, \hat{P}^{\text{norme}}) = 0$, tandis que $\text{Cov}(\hat{P}^p, \hat{P}^{\text{norme}}) = V(\hat{P}^p)$ pour l'une quelconque des provinces de référence. Une approximation est obtenue en négligeant les deux derniers termes de (2.3) :

$$V(\hat{E}^p) \approx \left(\frac{C^{\text{norme}}}{P^{\text{norme}}} \right)^2 V(\hat{P}^p) \quad (2.4)$$

En utilisant les données de la CVD de 2001, on peut vérifier que l'écart-type du paiement de péréquation dérive de (2.4) diffère de celui dérivé de (2.3) par au plus 7 %, sauf pour deux provinces bénéficiaires : Terre-Neuve-et-Labrador où l'approximation sous-estime l'écart-type de 11 %, et le Québec où l'approximation surestime l'écart-type de 12 %.

Comme nous pouvons le constater d'après l'équation (2.4), une répartition de l'échantillon qui produit des variances égales pour les estimations de la population des provinces bénéficiaires produit des variances égales pour les estimations des paiements de péréquation des provinces

bénéficiaires. Toutefois, le fait d'avoir des CV égaux pour les estimations de la population des provinces bénéficiaires ne garantit pas des CV égaux pour l'estimation des paiements de péréquation des provinces bénéficiaires, puisque, d'après l'équation (2.2), E_p^n n'est pas directement proportionnel à P_p^p parce que K_p^p n'est pas nul. Le fait d'avoir des CV égaux pour les estimations de la population des provinces bénéficiaires demeure un objectif valable, puisqu'il assure des intervalles de confiance de longueur égale pour le paiement de péréquation par personne. En fait, dû à l'utilisation de l'approximation (2.4), si la situation observée en 2001 se reproduit en 2006, l'intervalle de confiance pour Terre-Neuve-et-Labrador sera 11 % trop court (c'est-à-dire que la précision pour le paiement de péréquation par personne sera inférieure à celle des autres provinces bénéficiaires), tandis que l'intervalle de confiance pour le Québec sera 12 % trop long (c'est-à-dire que la précision pour le paiement de péréquation par personne sera supérieure à celle des autres provinces bénéficiaires). En outre, si nous faisons abstraction de la migration interprovinciale, alors les estimations démographiques provinciales sont indépendantes et la variance du total des paiements de péréquation est réduite au minimum si et seulement si la variance de la population totale des provinces bénéficiaires est réduite au minimum.

Nous cherchons à obtenir une répartition de l'échantillon au niveau provincial qui produit des CV égaux pour l'estimation de la population de chaque province bénéficiaire (objectif IV), de manière à obtenir une précision égale pour le paiement de péréquation par personne. La plus grande partie de la variation des estimations démographiques est attribuable à la variation des estimations du sous-dénombrement. Si l'on fait abstraction de la contribution du surdénombrement à la variation de l'estimation démographique, alors il est facile de vérifier que l'erreur type du taux estimé de sous-dénombrement est égale au CV de l'estimation démographique. Les objectifs I et II peuvent alors être reformulés comme étant de réduire au minimum le CV de l'estimation nationale de la population et de produire des estimations démographiques provinciales dont le CV est égal, respectivement. La différence entre les objectifs III et I, et entre les objectifs IV et II, c'est que ceux-ci s'appliquent, dans le premier cas, aux provinces bénéficiaires et, dans le deuxième cas, à toutes les provinces. Dans ce qui suit, nous supposons effectivement que la variance des estimations du sous-dénombrement.

Les objectifs de la répartition provinciale de l'échantillon sont résumés dans le tableau 2.1.

La répartition infraprovinciale optimale est simplement donnée par la répartition de Neyman. La difficulté est de prévoir la variance dans des strates relativement petites ou, plus précisément, de prévoir les totaux (nombre de personnes omises au recensement, nombre de non-répondants dans l'échantillon de la CVD) dont dépend la variance. Pour chaque province, l'approche utilisée dans cet article consiste à prendre d'abord les valeurs nationales plus stables au niveau de la cellule (âge × sexe × état matrimonial) et à les mettre à l'échelle de sorte que les totaux correspondent aux valeurs provinciales pour chaque groupe d'âge, pour chaque sexe et pour chaque état matrimonial. Cet objectif rappelle celui de la méthode d'ajustement proportionnel itératif proposée par Deming et Stephan (1940) et utilisée également par Rao (1976). Deville et Samdal (1992) ont montré comment on peut utiliser le calage pour obtenir le même résultat. Dans le cas de la CVD, on aura recours au calage même s'il est impossible d'aligner les cellules dans une matrice à trois dimensions, étant donné que les groupes d'âge diffèrent pour chaque état matrimonial. La méthode itérative du quotient parfois ne converge pas en raison de l'impossibilité de respecter les contraintes. En énonçant le problème de calage comme dans Théberge (1999), on tient compte de la possibilité que les contraintes soient incohérentes et ceci ne cause pas de problèmes de convergence. En outre, l'utilisation de l'inverse de Moore-Penrose dans la solution permet aussi aux contraintes d'être linéairement dépendantes.

Dans la section qui suit, nous examinerons la relation entre les estimations démographiques et les paiements de péréquation. Comme nous le constaterons, le problème de répartition de l'échantillon exige de trouver un juste équilibre entre quatre objectifs. À la section 3, nous utilisons une formule de variance approximative qui s'appuie sur un effet du plan pour déterminer la répartition optimale qui découle de chacun des quatre objectifs. Nous déterminons empiriquement la valeur de l'effet du plan à la section 4. À la section 5, nous expliquons comment une répartition finale peut établir un juste équilibre entre les répartitions individuelles pour des objectifs distincts. Enfin, la section 6 porte sur la répartition infraprovinciale. Nous n'examinons pas dans cet article la répartition de l'échantillon pour les trois territoires.

2. Incidence des estimations démographiques sur les paiements de péréquation

Statistique Canada est chargé de produire des estimations démographiques. Ces estimations démographiques trouvent une utilisation importante dans le calcul des paiements de péréquation effectué par le ministère fédéral des finances. Bien que Statistique Canada ne soit pas directement

ou

et

concerné par la formule servant à calculer les paiements de péréquation, il est utile de déterminer l'effet de la précision des estimations démographiques sur la précision des paiements de péréquation. L'incidence de la répartition de l'échantillon sur la précision des estimations démographiques retient l'attention depuis longtemps; dans cet article, nous examinerons également l'incidence de la répartition de l'échantillon sur la précision des paiements de péréquation. On utilise la CVD pour mesurer le taux de personnes omises au recensement. Dans le passé, la répartition de l'échantillon visait à la fois à réduire au minimum la variance pour le taux national estimé de sous-dénombrement (objectif I) et à produire des variances égales pour les taux estimés de sous-dénombrement dans chaque province (objectif II). Deux autres objectifs seront ajoutés en examinant l'impact de la répartition de l'échantillon sur la précision des paiements de péréquation, avant tout lissage basé sur des moyennes mobiles.

La formule servant à calculer les paiements de péréquation, avant tout lissage basé sur des moyennes mobiles, est :

$$E_p^p = \sum_{j=1}^{33} R_{ij}^p \left(\frac{T_{ij}^p}{T_{norm}^p} - \frac{P^p}{T^p} \right) \quad (2.1)$$

où E_p^p est le paiement de péréquation pour la province bénéficiaire p (au moment d'écrire ces lignes, toutes les provinces sauf l'Ontario et l'Alberta), R_{ij}^p représente les recettes totales (toutes les provinces) provenant de la source de recettes j , T_{ij}^p est l'assiette fiscale totale pour la source de recettes j , T_{norm}^p est l'assiette fiscale des provinces de référence servant à établir la norme (toutes les provinces sauf les provinces de l'Atlantique et l'Alberta) pour la source de recettes j , P^p est la population des provinces de référence, T_{ij}^p est l'assiette fiscale de la province bénéficiaire p pour la source de recettes j , et P^p est la population de la province bénéficiaire p .

Pour mesurer l'incidence des estimations démographiques sur les paiements de péréquation, nous réécrivons l'équation (2.1) comme suit :

$$E_p^p = \left(\frac{P^p}{P^{norm}} \right) C^{norm} - K^p, \quad (2.2)$$

$$C^{norm} = \sum_{j=1}^{33} R_{ij}^p \frac{T_{ij}^p}{T_{norm}^p}$$

$$K^p = \sum_{j=1}^{33} R_{ij}^p \frac{T_{ij}^p}{T^p}$$

Répartition de l'échantillon de la contre-vérification des dossiers de 2006

Alain Théberge

Résumé

La répartition d'un échantillon peut être optimisée en fonction de divers objectifs. Lorsqu'il y a plus d'un objectif, on doit choisir une répartition qui équilibre ces objectifs. Traditionnellement, la Contre-vérification des dossiers a établi cet équilibre en consacrant une fraction de l'échantillon à chacun des objectifs (par exemple, les deux tiers de l'échantillon sont répartis de manière à obtenir de bonnes estimations provinciales, tandis qu'un tiers est réparti de manière à obtenir une bonne estimation nationale). Cet article suggère une méthode qui consiste à choisir le maximum de deux ou plusieurs répartitions des estimations démographiques sur les paiements de l'échantillon de l'échantillon de la Contre-vérification des dossiers. La répartition intraprovinciale de la Contre-vérification des dossiers exige un lissage de paramètres définis au niveau des strates. Cet article montre comment le calage peut servir à ce lissage. Le problème de calage et sa solution n'existent pas l'existence d'une solution aux contraintes de calage. Ceci évite des problèmes de convergence rencontrés par des méthodes connexes telles l'ajustement proportionnel itératif (raking).

Mots clés : Ajustement proportionnel itératif; calage; contre-vérification des dossiers; lissage; répartition de l'échantillon.

1. Introduction

Le Recensement de la population du Canada est réalisé tous les cinq ans; il l'a été la dernière fois en 2001. La Contre-vérification des dossiers (CVD) vise à mesurer le surdénombrément du recensement et une partie du échantillon de la CVD, qui sera menée en 2006, on espère que la plus grande partie du échantillon de la CVD de la couverture sont utilisées conjointement avec les chiffres du recensement pour produire des estimations démographiques. Les estimations démographiques servent notamment au calcul, par le ministère fédéral des finances, des paiements de péréquation du gouvernement canadien aux gouvernements provinciaux. Habituellement, on procède à la répartition de l'échantillon de la CVD aux provinces en tâchant de trouver un juste milieu entre la nécessité de produire une estimation nationale de qualité du taux de personnes omises au recensement et la nécessité de produire des estimations provinciales de bonne qualité de ces taux, afin de produire les estimations démographiques de Statistique Canada. On espère que cette approche répondrait également aux besoins de produire des estimations de bonne qualité des paiements de péréquation (il s'agit d'estimations dans la mesure où ils sont fondés sur les estimations démographiques), mais cela n'a jamais été vérifié. Les paiements de péréquation sont effectués par le gouvernement fédéral

canadien aux provinces moins prospères. Dans le présent article, nous examinons l'effet de la répartition provinciale de l'échantillon sur la qualité des estimations des paiements de péréquation. Si la variance d'une variable d'intérêt est la même dans chaque province, alors on obtient une répartition optimale pour une estimation nationale dont la variance est minimale si la taille de l'échantillon est proportionnelle à la taille de la base de sondage pour la province p , N_p^d . Une répartition qui donne des estimations provinciales de variance égale est une répartition où la taille de l'échantillon est constante (proportionnelle à N_p^d). Une façon de mettre en équilibre les deux besoins souvent utilisée consiste à rendre la taille de l'échantillon proportionnelle à $N_p^{1/2}$. Aux fins de la CVD, on a utilisé dans le passé une méthode différente pour réaliser cet équilibre, selon laquelle une partie de l'échantillon est répartie de manière à produire des estimations provinciales de variance égale et l'autre partie est répartie de manière à produire une estimation nationale dont la variance est minimale. Habituellement, environ les deux tiers de l'échantillon sont répartis de manière à produire des estimations provinciales de variance égale. Nous proposons dans le présent article une nouvelle méthode de répartition provinciale qui établit un juste équilibre entre deux objectifs ou plus. Elle consiste à calculer une répartition distincte pour chaque objectif, chaque répartition portant peut-être sur une taille totale d'échantillon différente; on obtient la répartition finale, qui devrait répondre à tous les objectifs, en prenant pour chaque province la taille d'échantillon maximale sur chacune des répartitions.

des paramètres du modèle (Zanutto 1998, partie 1, annexe A). À l'étape 2, le deuxième cas n'est pas nécessaire pour maximiser la vraisemblance, mais il est inclus pour obtenir des prédictions pour les cellules de non-répondants non échantillonnées (c'est-à-dire, $i \notin S$, $r = 0$).

Bibliographie

- Beil, W.R., et Otto, M.C. (1994). Investigation of a model-based approach to estimation under sampling for nonresponse in the decennial census. Article non-publié présenté à la Joint Statistical Meetings, Toronto.
- Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B*, Methodological, 25, 220-233.
- Brakstone, G.J., et Rao, J.N.K. (1976). Raking ratio estimators. *Techniques d'enquête*, 2, 63-69.
- Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., et Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86, 68-78.
- Cox, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.
- Darroch, J.N., et Raftery, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43, 1470-1480.
- Dempster, A.P., Laird, N.M., et Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-22.
- Fuller, W.A., Isaki, C.T., et Tsay, J.H. (1994). Design and estimation for samples of census nonresponse. Dans *Proceedings of the Bureau of the Census Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, 289-305.
- Gelman, A., Carlin, J.B., Stern, H.S., et Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall Ltd.
- George, J.A., et Penny, R.N. (1987). Initial experience in implementing controlled rounding for confidentiality control. Dans *Proceedings of the Bureau of the Census Annual Research Conference*, Volume 3. Washington, DC: U.S. Bureau of the Census, 253-262.
- Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.
- Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Deuxième édition. New York: John Wiley & Sons, Inc.
- Meng, X.-L., et Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267-278.
- Oh, H.T., et Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. Dans *Incomplete Data in Sample Surveys* (Eds. W.G. Madrow, I. Olkin et D.B. Rubin). New York: Academic Press, 143-184.
- Purcell, N.J., et Kish, L. (1980). Postcensal estimates for local areas (or domains). *Revue Internationale de Statistique*, 48, 3-18.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Redfern, P. (1989). L'expérience européenne relative à l'utilisation des données administratives pour recenser la population : Questions d'ordre politique. *Techniques d'enquête*, 15, 85-103.
- Rubin, D.B., et Schenker, N. (1987). Interval estimation from multiply-imputed data: A case study using census agriculture industry codes. *Journal of Official Statistics*, 3, 375-387.
- Schafar, J.L. (1995). Model-based imputation of census short-form items. Dans *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: Bureau of the Census, 267-299.
- Schindler, E. (1993). Sampling for the count; sampling for non-mail returns. Rapport non-publié. U.S. Bureau of the Census.
- U.S. Bureau of the Census (1997a). Census 2000 operational plan. Washington, DC.
- U.S. Bureau of the Census (1997b). Report to Congress – the plan for Census 2000. Washington, DC.
- Wilkinson, G.N., et Rogers, C.E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, 22, 392-399.
- Zanutto, E. (1998). *Imputation for Unit Nonresponse: Modeling Sampled Nonresponse Follow-up, Administrative Records, and Matched Substitutes*. Thèse de maîtrise, Harvard University, Cambridge, Massachusetts.
- Zanutto, E., et Zaslavsky, A.M. (1995a). A model for imputing nonsample households with sampled nonresponse follow-up. Dans *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- Zanutto, E., et Zaslavsky, A. M. (1995b). Models for imputing nonsample households with sampled nonresponse follow-up. Dans *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: U.S. Bureau of the Census, 673-686.
- Zanutto, E., et Zaslavsky, A.M. (2002). Using administrative records to improve small area estimation: An example from the U.S. Decennial Census. *Journal of Official Statistics*, 18, 559-576.
- Zaslavsky, A.M. (1988). Redressement des estimations régionales par une pondération des ménages. *Techniques d'enquête*, 14, 281-305.
- Zaslavsky, A.M. (1993). Combining census, dual-system, and evaluation study data to estimate population shares. *Journal of the American Statistical Association*, 88, 1092-1105.
- Zaslavsky, A.M. (2004). Representing the Census undercount by multiple imputation of households. Dans *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives* (Eds. A. Gelman et X.-L. Meng). West Sussex, England: John Wiley & Sons, Inc. 129-140.
- Zhang, L.-C., et Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society, Series B*, 66, 479-496.

statistiques minimales suffisantes du modèle sont suffi-

(sammet proches.)

Notre algorithme tire parti du fait que les observations partiellement classifiées ne contribuent à la vraisemblance

que par la voie du nombre total de ménages dans chaque lots. Par conséquent, pour maximiser cette part de la vraisemblance, nous devons nous assurer que le nombre ajusté de non-répondants dans chaque lots est égal au nombre observé, ce qui est automatique, parce que l'interaction $\text{lots} \times \text{réponse}$ est toujours incluse dans notre

modèle.

L'algorithme API propose une méthode aux observations entièrement classifiées au moyen d'un algorithme API ordinaire, en ne tenant pas compte des observations partiellement croisées. Dans le cas du plan d'échantillonnage des flots, cela signifie que le modèle est ajusté en utilisant la partie entièrement observée du tableau $\text{flot} \times \text{type} \times \text{réponse}$ en utilisant un algorithme API ordinaire, en ne tenant pas

obtenir les prédictions pour les cellules partiellement croisées en appliquant à ces cellules les mêmes proportions croisées en appliquant à ces cellules partiellement croisées les prédictions pour les cellules partiellement croisées.

de sorte que le nombre ajusté de non-répondants dans chaque îlot soit égal au nombre observé. Dans le cas du plan d'échantillonnage des logements, nous utilisons le même

[illegible]

L'échantillon de suivi comme étant analogues aux lots hors de l'échantillon dans le cas de l'échantillonnage des lots.

Ceci donne des prédictions pour les ménages non-répondants dans les îlots sans ménages non-répondants dans les îlots de suivi. Nous obtenons les prédictions pour les ménages non-répondants dans les îlots comptant un ou plusieurs ménages non-répondants dans l'échantillon de suivi en appliquant la répartition prévue des types de ménage entre les ménages non-répondants échantillonnés dans chacun de ces îlots aux ménages non-répondants échantillonnés correspondants dans ces îlots. Pour plus de détails sur le cas de l'échantillonnage des logements, voir

Zanotto et al. Zaslavsky (2002).

Nous illustrons maintenant l'algorithme API pour l'échantillonnage des flots sous un modèle de Poisson tel

que avec $(\mathbf{g}(\mathbf{u}))^{\mathbf{d}_i} = (\mathbf{g}(\mathbf{u}))^{\mathbf{d}_i} \mathbf{z} = \mathbf{g}(\mathbf{u})^{\mathbf{d}_i} \mathbf{u}$ no \mathbf{d}_i représente une situation de réponse r , et Z est la matrice de plan d'expérience correspondant à l'expression du modèle $X^* * \mathbf{d} + \mathbf{d} * \mathbf{I} + X * \mathbf{I}$ II s'agit d'une version simplifiée du modèle donné par (2) ne comportant ni un ni plusieurs

consiste à fixer

Ces étapes sont répétées jusqu'à ce que les estimations des statistiques minimales suffisantes pour le modèle, à l'exclusion de m^{i+r} pour $i \notin S$ et $r = 0$ (c'est-à-dire, m^{i+r} pour $i \in S$ et $i \notin S$, $r = 1, m^{ij}$ pour $i \in S, m^{ij}$ pour $i \notin S, m^{ij(0)}$ et $\sum_{i \in S} m^{ij(0)}$) soient suffisamment approchantes de leurs valeurs observées. Si nous dénotons l'étape à laquelle ceci a lieu par i^* , l'étape finale de cet algorithme

À chaque étape, les facteurs d'échelle sont fondés unique-

$$\text{Etape 3 : } \hat{m}_{t+1}^{f1} = \hat{m}_{t+\frac{2}{3}}^{f1} \frac{n_{t+1}^{f1}}{n_{t+\frac{2}{3}}^{f1}} = \hat{m}_{t+\frac{2}{3}}^{f0} \frac{\sum_{i \in \mathcal{S}} \hat{m}_{t+\frac{2}{3}}^{f0} i^{\frac{f1}{3}}}{\sum_{i \in \mathcal{S}} \hat{m}_{t+\frac{2}{3}}^{f0} i^{\frac{f0}{3}}}$$

$$\text{Étape 3 : } m_{i+1}^{t+1} = m_{i+1}^{t+3} = \frac{m_{i+1}^{t+1}}{2}$$

$$\left. \begin{array}{l} \text{Step 1 : } m_{t+\frac{1}{3}}^{jfr} = \left\{ m_{t+\frac{r}{3}}^{jfr} \right\} \\ \text{si } i \in S \text{ ou si } i \notin S, r = 0 \end{array} \right\} \left\{ m_{t+\frac{1}{3}}^{jfr} \right\}$$

les trois étapes qui suivent dans le cycle :

L'algorithme API pour ajuster ce modèle débute par les estimations initiales $\hat{m}_{ijr}^0 = 1$ pour tout i, j, r et contient

On représente l'ensemble S par un diagramme de Hasse, mais uniquement si $0 \in S$ et $1 \in S$, mais uniquement si $0 \in S$ et $1 \in S$. Nous observons n types de ménage. La classification croisée complète de S est donnée par un diagramme de Hasse, mais uniquement si $0 \in S$ et $1 \in S$.

Zanutto et Zaslavsky : Un modèle d'estimation et d'imputation des ménages du recensement non-répondants

produit des estimations dont l'erreur est beaucoup plus faible que les deux autres méthodes examinées pour certaines variables étudiées, et à peu près équivalentes pour d'autres. Ces conclusions tiennent pour le plan d'échantillonnage des îlots ainsi que celui des logements. L'un des avantages de notre approche est que les modèles peuvent être spécifiés de façon à n'imposer des contraintes que sur quelques tableaux de marge ou interactions des caractéristiques aux niveaux de détail géographique les plus fins, où les données sont peu nombreuses, tout en ajustant des distributions plus détaillées des caractéristiques à des niveaux d'aggrégation géographique plus élevés auxquels un grand nombre de données sont disponibles. Cette approche est en harmonie avec les pratiques habituelles concernant la diffusion des données du recensement, qui contiennent un nombre minimal de caractéristiques au niveau de l'îlot, mais des caractéristiques de plus en plus détaillées pour les plus grandes unités.

De nombreuses applications importantes des données du recensement comportent l'estimation de la population et de ses caractéristiques pour de petits domaines tels que les districts législatifs et les secteurs de planification des services sociaux (comme les écoles et les cliniques) et du développement commercial. Bien que ces domaines ne coïncident pas toujours avec les secteurs utilisés pour le calcul des estimations de recensement, le fait de contrôler les estimations du recensement de façon à ce qu'elles concordent avec des estimations sans biais à plusieurs niveaux de détail géographique accroît la probabilité que les estimations calculées pour des domaines pertinents pour l'élaboration des politiques créés en regroupant le tout ou certaines parties de ces secteurs seront également presque sans biais. Notre méthode donne des propriétés au niveau agrégé plus prévisible que des alternatives complexes comme la modélisation spatiale hiérarchique. Bien que cette dernière puisse produire des estimations dont l'erreur quadratique moyenne est plus faible aux niveaux les plus fins de détail géographique, l'ajustement de ce genre de modèle et la vérification de leur biais à divers niveaux d'aggrégation géographique nécessiteraient des mises au point locales de grande portée qui seraient vraisemblablement irréalistes dans les conditions de production d'un recensement.

Notre méthodologie est illustrée ici dans le contexte d'un échantillonnage pour le suivi des cas de non-réponses pour le recensement décennal des États-Unis, mais notre stratégie d'estimation et d'imputation peut être utilisée pour l'estimation et l'imputation pour de petits domaines dans le cadre de tout recensement ou enquête utilisant un échantillonnage SCNR où les populations présentent une structure hiérarchique. Nous pouvons aussi intégrer des enregistrements administratifs comme covariables afin de prédire les

caractéristiques des ménages non-répondants correspondants (Zanutto et Zaslavsky 2002). Dans un tel scénario, les données sur les ménages inclus dans l'échantillon SCNR pour lesquelles nous possédons des renseignements provenant à la fois du recensement et des enregistrements administratifs sont utilisées pour estimer les écarts systématiques entre les deux sources de renseignements. Sous les mêmes modèles, nous imputons les caractéristiques des ménages non-répondants non échantillonnés. L'utilisation des enregistrements administratifs dans cette approche de modélisation peut améliorer l'exactitude des estimations sur petits domaines (niveau de l'îlot).

La discussion de l'échantillonnage dans le contexte du recensement des États-Unis s'avère politiquement litigieuse, mais il n'en reste pas moins qu'à long terme, il est probable qu'une forme ou l'autre d'estimation sera utilisée pour les non-répondants. Les possibilités pourraient être encore plus grandes dans les pays où les estimations démographiques reposent déjà sur une utilisation importante des dossiers administratifs (Redfern 1989). Des méthodes telles que celles décrites ici permettant de combiner l'information provenant de plusieurs sources de données tout en reflétant la diversité locale seront des éléments essentiels de ce genre d'efforts.

Annexe

Ajustement proportionnel itératif avec données partiellement croisées

Une approche type de l'ajustement de modèles logarithmiques à des données partiellement croisées consiste à utiliser un algorithme EM (Dempster, Laird et Rubin 1977; Little et Rubin 2002, chapitre 8) dans lequel, par étapes alternées, 1) les dénominateurs prévus sont imputés sous les conditions du modèle et 2) le modèle est rajusté aux données observées et imputées par la technique de l'ajustement proportionnel itératif (API) (Darroch et Raftery 1972) pour des modèles sans solutions analytiques. Dans la modification ECM plus efficace de cet algorithme, un seul cycle de l'algorithme API est réalisé à chaque étape (Meng et Rubin 1993).

Dans le cas de notre application, nous avons développé un algorithme API plus rapide que les algorithmes EM et ECM pour nos modèles, qui comprend toujours une interaction îlot \times réponse et ne comprend aucune interaction îlot \times type \times réponse. Nous avons constaté que notre algorithme API modifié converge après environ la moitié ou les deux tiers du nombre de cycles qu'exige l'algorithme EMC, en demandant moins de calculs à chaque étape (Zanutto 1998, partie I, annexe A). (Nous déclarons qu'il y a convergence quand les valeurs prévues et observées des

La RMW MSE est nettement plus faible pour la méthode du ratio stratifiée et le modèle loglinéaire que pour la méthode du ratio non stratifiée pour la plupart des caractéristiques des ménages au niveau de l'ilot et du secteur. Par conséquent, nous limitons la suite de la discussion à la comparaison des deux premières méthodes.

Les différences les plus importantes se dégagent pour les catégories de mode d'occupation du logement aux niveaux de l'ilot et du secteur. Dans chaque DR, les estimations des catégories de mode d'occupation aux niveaux de l'ilot et du secteur produites par le modèle loglinéaire ont une RMW MSE beaucoup plus faible que celles obtenues par la méthode du ratio stratifiée, principalement parce que le modèle loglinéaire donne lieu à un biais nettement plus petit (RMSB). Les écarts-types (RMWV) sont un peu plus grands pour le modèle loglinéaire sous échantillonnage des logements, mais à peu près égaux pour les deux méthodes sous l'échantillonnage des ilots. Le modèle loglinéaire produit un biais plus faible pour les catégories de mode d'occupation au niveau du secteur, parce que le mode d'occupation du logement est inclus dans le modèle à titre d'effet au niveau du secteur, x_i . La stratification en fonction de la race dans la méthode du ratio réduit la RMW MSE pour les catégories de race au niveau de l'ilot, mais les deux méthodes donnent une RMW MSE comparable pour les catégories de race aux niveaux du secteur et du DR. La méthode du ratio stratifiée perd son avantage par rapport au modèle loglinéaire au niveau du secteur, parce qu'elle n'utilise aucune information au niveau du secteur. Dans l'ensemble, les deux méthodes produisent des estimations dont la RMW MSE est comparable à tous les niveaux de détail géographique pour les catégories de taille.

La signification statistique (dans les conditions des simulations) des écarts entre les RMW MSE des diverses méthodes a été évaluée au moyen de tests t . Presque tous les écarts observés aux niveaux de l'ilot et du secteur, à l'exclusion de la catégorie des logements inoccupés, ont une valeur $p \leq 0,001$ pour le test bilatéral et, par conséquent, ne peuvent être attribués à une erreur de simulation.

5. Évaluation et prédiction de l'erreur de modélisation

Faute d'espace, nous nous limitons ici à résumer brièvement les méthodes d'estimation de l'erreur quadratique moyenne des estimations ajustées en utilisant les données d'échantillon. Les méthodes et les résultats peuvent être obtenus en s'adressant au premier auteur.

Pour commencer, nous avons élaboré des approximations analytiques qui prédisent l'effet de la variation du taux d'échantillonnage sur l'exactitude de nos estimations sans mentaires pour chaque taux. Ces approximations peuvent

être utiles pour l'établissement du plan d'échantillonnage. Nous produisons une approximation de la RMW MSE des estimations aux niveaux de l'ilot, du secteur et du DR pour un nouveau taux d'échantillonnage, selon le plan d'échantillonnage des ilots ainsi que celui des logements, sachant que des résultats de simulation pour un premier taux d'échantillonnage existent déjà, en combinant les estimations du biais et de la variance au taux d'échantillonnage courant au moyen de deux facteurs de rééchantillonnage. Le premier reflète la nouvelle proportion de logements sur lesquels doivent porter l'estimation sous le nouveau taux d'échantillonnage, qui a une incidence sur le biais et la variance des estimations combinées. L'autre facteur reflète l'effet du taux d'échantillonnage sur la variance des estimations. Les simulations ont démontré l'exactitude des prédictions de la RMW MSE lorsqu'on utilise ces approximations, sauf pour certaines extrapolations extrêmes.

À l'aide de ces résultats, nous avons mis au point une procédure de contrevalidation pour faciliter l'obtention des estimations intra-échantillon de la RMW MSE à utiliser dans des conditions de production où les caractéristiques réelles des ménages non-répondants ne sont pas connues. Pour chaque secteur, l'échantillon de suivi est subdivisé aléatoirement en C groupes de contrevalidation (d'ilots pour l'échantillonnage des ilots et de ménages pour l'échantillonnage des logements). Les groupes de contrevalidation sont éliminés chacun à leur tour et le modèle est ajusté aux non-répondants dans les $C-1$ groupes de contrevalidation restants et aux répondants dans l'ensemble des C groupes. Nous pouvons alors estimer la RMW MSE dans les conditions du plan d'échantillonnage simulées par des contrevalidations et projeter cette estimation au taux d'échantillonnage courant, ou tout autre taux d'intérêt, en utilisant les approximations décrites au paragraphe précédent. Les simulations indiquent que cette procédure donne des estimations exactes de la RMW MSE aux niveaux de l'ilot et du DR, et une certaine surestimation au niveau du secteur. Cette méthode fournit aussi des estimations distinctes du biais et de la variance qui, d'après les simulations, sont très précises. Ces résultats sont très utiles pour évaluer l'adéquation du modèle, car un mauvais ajustement sera trahi par une composante d'erreur quadratique moyenne importante due au biais.

6. Conclusion

Aux sections précédentes, nous avons présenté une approche basée sur un modèle pour imputer les caractéristiques des ménages non-répondants à un recensement non-réponse. Dans les simulations, notre modèle loglinéaire

Cela nous donne des estimations convergentes de l'erreur lors de l'agrégation sur les unités géographiques, ce qui est approprié étant donné le caractère arbitraire des limites des unités (Zaslavsky 1993). Nous fondons nos mesures sur les erreurs de mesure relatives à la population totale de la région géographique ! plutôt qu'à la population dans la catégorie cible uniquement, parce que ce dernier dénominateur exagère l'importance des petites erreurs pour les ilots dans lesquels la catégorie n'apparaît que rarement ou jamais.

4.4 Résultats

Dans le cas de la simulation de l'échantillonnage SCNR selon un plan d'échantillonnage des ilots ainsi qu'un plan d'échantillonnage d'unités de logement, nous estimons le nombre de ménages possédant chaque caractéristique aux niveaux de l'îlot, du secteur et du DR au moyen de chacune des trois méthodes d'estimation. À la figure 1, les résultats

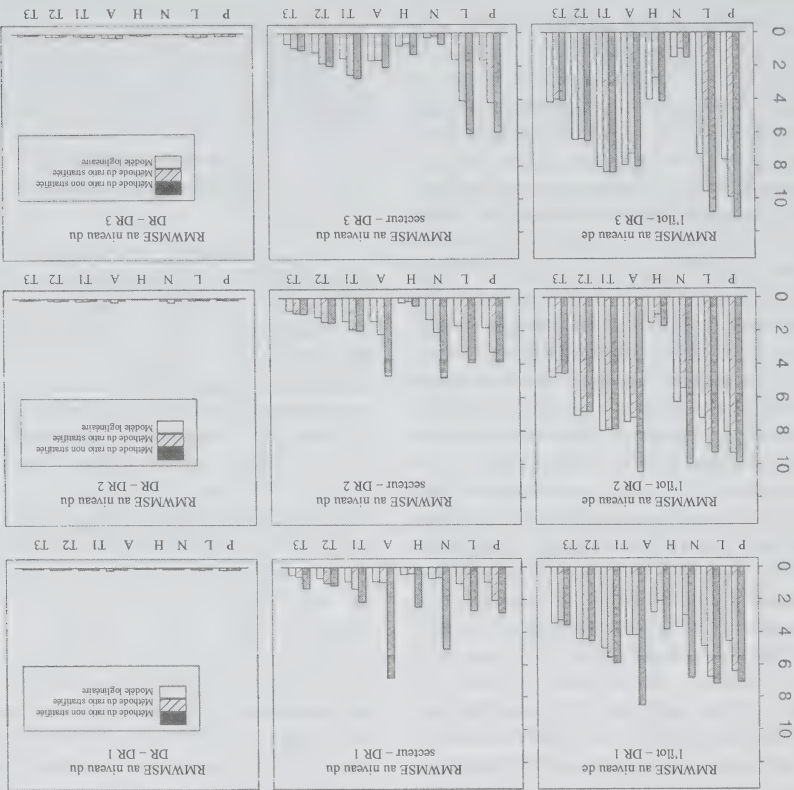


Figure 1. Estimations de la RMW MSE aux niveaux de l'îlot, du secteur et du DR pour chaque caractéristique des ménages, en utilisant le plan d'échantillonnage des logements pour les DR 1, 2 et 3, avec des échantillons simulés. (P = propriétaire, L = locataire, N = Noir, H = Hispanique, A = Autre race, T1 = Groupe de taille 1 (1 à 2 personnes), T2 = Groupe de taille 2 (3 à 4 personnes), T3 = Groupe de taille 3 (5 personnes et plus).

pour chaque méthode sont représentées par les diverses barres ombrées pour l'échantillonnage des logements. (Nous ne présentons pas les résultats pour l'échantillonnage des ilots, mais le profil des résultats est comparable, la RMW MSE étant environ 10 % plus élevée pour toutes les estimations.) Dans cette figure, chaque ligne de graphiques à barres donne la RMW MSE pour les estimations aux niveaux de l'îlot, du secteur et du DR, pour l'un des trois DR. Chaque groupe de trois barres représente la RMW MSE pour les estimations du nombre total de ménages selon chaque catégorie de mode d'occupation du logement, chaque catégorie de taille de ménage et chaque catégorie de race en utilisant chacune des trois méthodes. Comme les trois méthodes s'appuient sur le même modèle de régression logistique pour prédire le nombre de logements non-répondants non échantillonnés inoccupés dans chaque bloc, nous avons omis la catégorie des logements inoccupés dans les graphiques.

(en moyenne, 548 ménages par secteur dans le DR 2 et 918 ménages par secteur dans le DR 3).

Tableau 1

Caractéristiques des districts de recensement utilisés dans les simulations

	DR1	DR2	DR3
Ménages	1 12 966	169 321	149 567
Îlots	4 907	15 470	8 167
Pseudo-secteurs	94	309	163
Noir non-hispanique	14,4 %	28,5 %	1,3 %
Hispanique	6,1 %	1,0 %	6,6 %
Autre	73,5 %	59,4 %	81,5 %
Propriétaire	63,8 %	59,5 %	52,6 %
Locataire	30,2 %	29,4 %	36,7 %
Logement inoccupé	6,0 %	11,1 %	10,7 %
Taille 1 (1 à 2 personnes)	50,4 %	46,9 %	55,2 %
Taille 2 (3 à 4 personnes)	31,6 %	31,6 %	26,2 %
Taille 3 (5 personnes et plus)	12,0 %	10,4 %	7,9 %
Taux de réponse	72,6 %	65,3 %	56,7 %

4.3 Mesures du biais, de la variance et de l'erreur quadratique moyenne

Les fonctions de perte que nous utilisons pour nos évaluations sont fondées sur l'erreur relative pour la catégorie de ménage j (un type ou une combinaison de types) dans l'unité géographique i (un îlot ou un groupe d'îlots) :

$$(3) \quad d_{jfs}^{i+} = \frac{Y_{jfs}^{i+} - Y_{ij}}{Y_{ij}}$$

où Y_{ij} est le nombre réel de ménages dans la catégorie j dans l'unité géographique i , Y_{jfs}^{i+} est le nombre correspondant de ménages estimés d'après l'échantillon s (y compris ceux observés dans l'échantillon et ceux estimés par le modèle), et Y_{i+} est le nombre total de ménages dans l'unité géographique i .

Nous résumons le biais dans les dénombrements estimés pour la catégorie j et un niveau de détail géographique (îlot, secteur, DR) au moyen de la racine carrée du biais quadratique moyen pondéré (RMWSB pour *Root Mean Weighted Squared Bias*) :

$$(4) \quad \text{RMWSB}_j^2 = \frac{\sum_i Y_{i+} \left\{ \left(\frac{1}{S} \sum_s d_{jfs}^s \right)^2 - \frac{1}{S(S-1)} \left(\sum_s d_{jfs}^s \right)^2 \right\}}{\sum_i Y_{i+}}$$

où S est le nombre d'échantillons tirés et $i = 1, \dots, I$ où I est le nombre d'unités géographiques. Le deuxième terme

du numérateur élimine un biais dû à la finitude de la simulation. Sous l'angle du plan de sondage, nous considérons la composition de chaque secteur comme une quantité fixe et seul l'échantillonnage est aléatoire. Alors on définit le biais comme la différence moyenne, sur tous les échantillons possibles, entre la valeur réelle pour un secteur donné et la valeur estimée correspondante; il s'agit essentiellement de l'erreur de modèle pour le secteur. Ce genre d'erreur est inévitable puisque la composition des non-répondants ne peut être entièrement prédite dans tout îlot. Un type plus grave de biais correspondrait à une erreur systématique dans les estimations pour un ensemble d'îlots de composition comparable. Nous n'avons pas recherché tous les types possibles de biais au sens susmentionné, mais la spécification du modèle nous protège contre ce biais aux niveaux élevés d'agrégation, parce que les estimations d'après le modèle sont contrainues de concorder (approximativement) avec des estimations sans biais pour les secteurs et les DR.

À titre de mesure de l'erreur globale, nous calculons la racine carrée de l'erreur quadratique moyenne pondérée (RMWMSE pour *Root Mean Weighted Mean Squared Error*) pour chaque catégorie de ménage j , qui est donnée par

$$(5) \quad \text{RMWMSE}_j^2 = \frac{\sum_i Y_{i+} \left(\frac{1}{S} \sum_s d_{jfs}^s \right)^2}{\sum_i Y_{i+}}$$

où Y_{ij} , Y_{jfs}^{i+} , Y_{i+} , i et S sont définis de la même façon que précédemment. (Les deux « moyennes » sont des moyennes sur les unités géographiques i et sur les échantillons s .) Nous obtenons une mesure de l'écart-type des estimations pour la catégorie de ménage j en calculant la racine carrée de la variance moyenne pondérée (RMWV pour *Root Mean Weighted Variance*) :

$$(6) \quad \text{RMWV}_j^2 = \frac{\sum_i Y_{i+} \left\{ \frac{1}{S-1} \left(\sum_s d_{jfs}^s \right)^2 - \left(\frac{1}{S} \sum_s d_{jfs}^s \right)^2 \right\}}{\sum_i Y_{i+}} = \text{RMWMSE}_j^2 - \text{RMWSB}_j^2.$$

Il convient de souligner que ces mesures de l'erreur quadratique moyenne, du biais et de l'écart-type sont toutes des estimations se rapportant aux échantillons SCNR répétés à partir de la population finie d'îlots. Ces fonctions de perte peuvent être appliquées à divers niveaux de détail géographique, ce qui reflète le fait que l'utilisation principale des estimations au niveau de l'îlot est l'agrégation pour former des estimations de niveau géographique plus élevée. En tenant compte de cet aspect, nous avons également choisi ces mesures parce qu'elles pondèrent les erreurs par la taille de l'unité géographique.

proportionnel itératif, nous lisons les données en ajoutant à chaque îlot un ménage répondant hypothétique (« pseudo-domnées »). Ce ménage est réparti entre les 18 types de ménages correspondants aux logements occupés conformément aux proportions globales de ménages répondants dans le DR. Les estimations obtenues en utilisant cinq ménages pour le lissage étaient à peu près de la même précision que celles obtenues avec un ménage, et un lissage plus agressif (par ajout de 10, 15, 20 ou 25 ménages par îlot) augmentaient les erreurs d'estimation. En outre, même si l'ajout d'une petite fraction uniquement d'un ménage à chaque îlot suffit à assurer que le modèle puisse être ajusté à chaque cas, l'utilisation de moins d'un ménage par îlot ralentit considérablement la convergence et augmente légèrement l'erreur dans les estimations.

Les trois méthodes d'estimation s'appuient sur le même modèle de régression logistique pour les logements inoccupés. Pour chaque îlot, les covariables sont le taux de non-réponse par la poste, les pourcentages de ménages répondants qui sont (séparément) locataires, occupants d'un appartement et membres d'une race minoritaire (Noirs ou Hispaniques), la valeur moyenne des logements occupés par leur propriétaire, le loyer mensuel moyen pour les logements locaux, des variables indicatrices pour chacun des secteurs et les interactions entre le pourcentage de locataires répondants et le loyer mensuel moyen, le pourcentage de locataires répondants et le carré du loyer mensuel moyen (centré sur la moyenne), le pourcentage de propriétaires répondants et la valeur moyenne des logements, et le pourcentage de propriétaires répondants et le carré de la valeur moyenne des logements (centré sur la moyenne). Afin d'éviter les problèmes de calcul causé par les îlots ne contenant pas de ménages non-répondants correspondants à des logements inoccupés, un ménage non-répondant hypothétique est ajouté à chaque îlot et réparti entre les logements inoccupés et occupés conformément à leur proportion dans les ménages non-répondants échantillonnés dans le DR.

4.2 Données

Nous utilisons les données du questionnaire abrégé du Recensement de 1990 pour trois DR dont les caractéristiques sont décrites au tableau 1. La race d'un ménage est déterminée par celle dont la prévalence est la plus élevée dans le ménage, habituellement la seule qui existe (98 % des ménages). Dans le DR 1, nous avons fusionné les groupes d'îlots consécutifs (et par conséquent contigus) (groupes d'îlots contigus) en 94 secteurs contenant, en moyenne, 52 îlots et 1 100 ménages. Pour les DR 2 et 3, nous ne disposons pas de renseignements sur les groupes d'îlots, si bien que nous avons formé des secteurs en groupant des îlots consécutifs en grappes contenant en moyenne 50 îlots

bas) et la méthode logarithmique sont inférieurs à 0,05, sauf quand la différence entre les RMSE estimées est très faible, auquel cas le coefficient de variation est très grand.

Nous comparons les propriétés de notre modèle à deux méthodes d'estimation de rechange, sous échantillonnage des unités de logement ainsi que des îlots. Chaque méthode commence par l'ajustement d'un modèle de régression logistique afin d'estimer, pour chaque îlot, le nombre de ménages non-répondants qui correspondent à des logements inoccupés. La première méthode, c'est-à-dire la « méthode du ratio non stratifiée », consiste à imputer des ménages dans chaque îlot pour remplacer les ménages non-répondants non échantillonnés proportionnellement à la répartition des types de ménage entre les ménages non-répondants dans l'échantillon de suivi pour le DR complet. La deuxième option, c'est-à-dire la « méthode du ratio stratifiée », est une variante de celle de FIT. Nous commençons par former des strates d'environ 82 îlots d'après la composition raciale de ces derniers, comme l'on décrit FIT. (Nous utilisons les données sur les répondants ainsi que les non-répondants pour former les strates, en supposant, comme FIT, que des renseignements semblables pourraient être tirés de dossiers administratifs. La stratification fondée uniquement sur l'information sur les répondants a donné des résultats comparables.) Puis, dans chaque strate, nous imputons des ménages non-répondants non échantillonnés aux divers types de ménage dans les logements occupés proportionnellement à la fréquence du type concerné dans l'échantillon de suivi pour la strate en question.

Nous simulons chaque méthode d'estimation en utilisant un taux d'échantillonnage SCNR de 30 %. Dans chaque strate, nous simulons cet échantillonnage en sélectionnant un échantillon aléatoire simple à 30 % d'îlots dans le cas du plan d'échantillonnage des îlots et un échantillon aléatoire simple à 30 % de ménages non-répondants dans chaque strate dans le cas du plan d'échantillonnage des logements. Nous supposons que les caractéristiques des ménages non-répondants dans ces échantillons sont connues (c'est-à-dire à la suite des opérations de suivi). Aussi bien pour la méthode du modèle logarithmique que pour la méthode du ratio stratifiée, nous avons sélectionné un échantillon à 30 % d'îlots ou de ménages non-répondants par échantillonnage aléatoire simple sans remise dans que région.

Nous considérons plusieurs formulations du modèle logarithmique. Tant pour l'échantillonnage des îlots que pour celui des logements, d'après les critères décrits à la section 4.3, le meilleur modèle est celui qui utilise $x_1 = \text{taille} * \text{race} * \text{mode d'occupation} + \text{taille} * \text{race} * \text{mode d'occupation}$, $x_2 = \text{race} * \text{mode d'occupation} + \text{taille} * \text{race} * \text{mode d'occupation}$, $x_3 = \text{race} * \text{taille}$, $x_4 = \text{mode d'occupation}$. Ce modèle est celui que nous avons utilisé dans les simulations.

Pour être certain que le modèle puisse être ajusté à chaque cas et pour accélérer la convergence de l'ajustement

4. Simulations

4.1 Vue d'ensemble

Notre étude en simulation a pour but d'évaluer le biais, la variance et l'erreur quadratique moyenne (EQM) des estimations des agrégats démographiques étudiés (comme le nombre de ménages selon la race, la taille et le mode d'occupation du logement) à divers niveaux de détail géographique, en utilisant les compositions estimées des ménages pour les ménages non-répondants non compris dans l'échantillon SCNR. Les évaluations analytiques sont inévitables étant donné la complexité des modèles et du plan d'échantillonnage, la relation de dépendance entre les propriétés du modèle et la répartition géographique réelle des types de ménage, ainsi que le nombre de variantes du modèle qui pourraient être examinées.

Nous avons utilisé des données au niveau de l'îlot du recensement décennal des États-Unis de 1990 provenant de trois districts de recensement (DR); ces îlots représentent simulations sont semblables à celles décrites par Schindler (1993) ou par FTT.

Les étapes de la simulation sont les suivantes :

1. Échantillonner les îlots ou les logements non-répondants conformément au plan d'échantillonnage SCNR.
2. Ajuster un modèle de régression logistique pour les ménages des logements inoccupés aux caractéristiques des ménages répondants et à celles des ménages non-répondants échantillonnés.
3. Calculer le nombre prévu de ménages non-répondants correspondant à un logement inoccupé pour chaque îlot.
4. Ajuster un modèle pour les types de ménages parmi les logements occupés en utilisant les caractéristiques des ménages répondants et des ménages non-répondants échantillonnés.
5. Calculer, pour chaque îlot, le nombre prévu de ménages non-répondants non échantillonnés pour chaque type de ménage occupés.
6. Calculer les agrégats d'intérêt d'après les dénombrements prévus et les comparer aux valeurs réelles au moyen de fonctions de perte.

Lors de l'exécution de nos simulations, répéter ces étapes 30 fois a produit des estimations de la racine carrée de l'erreur quadratique moyenne (RMSE pour Root mean square error) (définie à la section 4.3) d'une précision suffisante pour évaluer les propriétés de notre modèle comparativement à des modèles de rechange. Plus précisément, les coefficients de variation estimés des différences estimées de RMSE pour la méthode par le ratio stratifiée (décrite plus

Dans le cas de certains ensembles de données, il se peut que certains paramètres ne puissent être estimés parce que les estimations de vraisemblance se situent sur la frontière de l'espace du paramètre (infinité dans le cas de l'échelle logarithmique, ce qui donne une valeur nulle sur l'échelle de dénombrement) ou parce qu'il n'existe aucun renseignement pour le paramètre. Adapter la spécification du modèle à chaque domaine d'estimation pour éliminer les paramètres qui ne peuvent être estimés est irréaliste dans le contexte de production d'un recensement.

Grâce à l'introduction d'une petite quantité d'information a priori, il est possible d'assurer que tous les paramètres puissent être estimés. Pour cela, nous annexons aux données dont nous disposons pour chaque secteur une petite quantité de « pseudos-données » dont les proportions selon le type sont égales à celles observées pour une région environnante (le DR dans nos simulations), en ajoutant ces dénombrements au tableau de données avant d'ajuster le modèle. Cette étape correspond à l'application d'une analyse bayésienne empirique à des données multinomiales de loi $f(n_1, \dots, n_H | p_1, \dots, p_H) \propto \prod_{h=1}^H p_h^{n_h}$, où n_1, \dots, n_H sont les nombres observés de ménages de chaque type dans un îlot ou un secteur. Si $\{p_i\}$ suit une loi a priori conjointe de Dirichlet, $f(p_1, \dots, p_H) \propto \prod_{i=1}^H p_i^{\alpha_i - 1}$, $\alpha_i > 0$, la loi a posteriori résultante des p_i est une Dirichlet de paramètres $\alpha_i + x_i$ (Gelman, Carlin, Stern et Rubin 1995, page 76) et de mode a posteriori proportionnel aux paramètres. Donc, cette méthode bayésienne empirique équivaut à ajouter $\sum \alpha_i$ ménages au secteur, où les α_i de ces ménages sont du i^{e} type. Nous posons que les α_i sont proportionnels aux proportions observées pour chaque type de ménage dans une région environnante, de sorte que la distribution avec celle observée pour une région soit lissée par le mélange avec celle observée pour une région environnante, ce qui évite l'introduction d'un biais au niveau de la région plus grande. Cette spécification a priori induit une loi a priori sur les paramètres du modèle logarithmique. Voir Rubin et Schenker (1987), Zaslavsky (1988), ainsi qu'un exemple et des références historiques dans Clogg, Rubin, Schenker, Schultz et Weidman (1991) pour une utilisation semblable du lissage.

Quand les paramètres du modèle sont estimés, l'étape suivante consiste à calculer les dénombrements prévus de chaque type de ménage pour les ménages non-répondants qui ne font pas partie de l'échantillon SCNR. Au moyen de l'algorithme d'ajustement proportionnel itératif, nous obtenons automatiquement les prédictions pour les ménages non-répondants non échantillonnés en appliquant les mêmes proportions d'ajustement à la partie partiellement observée du tableau qu'à la partie entièrement observée, de sorte qu'aucun calcul supplémentaire n'est nécessaire (voir l'annexe).

au moyen du terme d'effet principal pour x_1) et (4) par ilot (pour les caractéristiques x_2 au moyen du terme $t * x_2$), et (5) caractéristiques moyennes des ménages pour l'ensemble des non-répondants (pour les caractéristiques x_3 au moyen du terme $r * x_3$) et (6) pour les répondants par secteur (pour les caractéristiques x_4 au moyen du terme $r * a * x_4$). Donc, ce modèle généralise le modèle d'indépendance ilot \times type utilisé par FTT et donne des résultats sans biais à des niveaux d'agrégation plus faibles, en supposant que les totaux de marges et les moyennes sont estimés sans biais d'après les données. L'estimation pour le secteur n'est pas exactement la même que l'estimation sans biais usuelle obtenue par estimation directe d'après l'échantillon SCNR, parce que le modèle fait concorder les valeurs de marge observées et ajustées pour les ménages compris dans l'échantillon. En effet, il existe un ajustement pour la covariance (régression) qui déplace l'agrégat de façon à tenir compte des différences observées entre les ménages répondants dans les ilots échantillonnés et les ménages répondants dans les ilots non échantillonnés, entre les ménages du plan d'échantillonnage des logements, entre les ménages répondants dans les ilots pour lesquels des ménages sont sélectionnés dans l'échantillon SCNR et ceux pour lesquels l'échantillon SCNR ne contient aucun ménage.

L'idée de modéliser les caractéristiques des ménages en utilisant des covariables de niveau peu détaillé au niveau de l'ilot et des covariables de niveau plus détaillé aux niveaux géographique plus agrégés est satisfaisable du point de vue conceptuel, mais non dans les détails, au modèle décrit dans Zaslavsky (2004). Pour une description de l'utilisation de poids logarithmiques pour faire concorder les estimations d'échantillon aux agrégats, voir Brackstone et Rao (1976), Oh et Scheuren (1983), et Zaslavsky (1988).

3.3 Estimation et lissage

Nous ajustons le modèle par estimation du maximum de vraisemblance sous échantillonnage de Poisson, ce qui équivaut à ajuster un modèle de régression logarithmique multinomiale. Le fait que les données ne forment pas un tableau ilot \times réponse \times type complet, parce que nous disposons des dénombrements par ilot, mais non des caractéristiques des ménages non-répondants non échantillonnés, complice l'ajustement du modèle. Dans le cas du plan d'échantillonnage des ilots, les renseignements sur les caractéristiques manquent pour tous les non-répondants dans certains ilots et, dans le cas du plan d'échantillonnage des logements, les renseignements sur les caractéristiques manquent pour certains non-répondants dans presque tous les ilots. Pour ajuster le modèle, nous utilisons un algorithme d'ajustement proportionnel itératif (API) modifié adapté aux données qui sont classées partiellement dans une partie de l'ensemble de données (voir l'annexe).

Toutes les interactions pourraient être incluses, sauf celles de la forme $r * t * x_1$, où x représente une expression du modèle en les variables qui définissent le type de ménage (c'est-à-dire telles que x_1, x_2, x_3 , ou x_4). Les interactions de cette forme dépendent des totaux de marge déterminés uniquement d'après les non-répondants dans les ilots non échantillonnés sous le plan d'échantillonnage des ilots et sont fondés sur un très petit échantillon sous le plan d'échantillonnage des unités de logement. Par conséquent, notre spécification du modèle exclut tous les effets $r * t * x_1$, qu'il est toujours impossible d'estimer (ou qui sont estimés médiocrement dans le cas du plan d'échantillonnage des ménages). Ce modèle généralise deux théories simples qui sont intégrées sous forme de sous-modèles. En premier lieu, s'il n'y a aucune différence entre les ilots (c'est-à-dire que les interactions logarithmiques $t * x_2$ et $a * x_4$ sont nulles), alors les ménages non-répondants dans chaque ilot sont imputés d'après la proportion globale de ménages non-répondants dans chacune des catégories x_3 dans l'échantillon SCNR, grâce à l'effet $r * x_3$. Autrement dit, les imputations sont faites en utilisant les mêmes proportions dans chaque ilot. En deuxième lieu, s'il n'existe aucune différence entre les répondants et les non-répondants (c'est-à-dire pas d'interaction $r * x_3$ ou $r * x_4$), alors les non-répondants sont imputés en utilisant les mêmes proportions que celles observées pour les répondants dans chaque ilot.

La formulation générale de notre modèle permet de tenir compte de nombreuses définitions du secteur et du type de ménage, et d'un choix nombreux d'expressions du modèle. Les secteurs devraient être définis de façon qu'ils soient suffisamment grands pour contenir des données adéquates pour l'estimation des interactions correspondantes, mais être aussi relativement homogènes. Par exemple, ils pourraient être définis par une combinaison de conglomérats géographiques et de stratifications selon les covariables au niveau de l'ilot (comme le pourcentage de minorités), afin d'obtenir des secteurs plus homogènes dont les différences pourraient être décrites par modélisation. La généralisation à plus de deux niveaux d'agrégation géographique dans le domaine d'estimation est également simple. Donc, nous pourrions par exemple ajouter l'interaction d'une autre expression du modèle x_5 avec une unité géographique de niveau compris entre celui du secteur et de l'ilot.

Lors de l'ajustement du modèle par la méthode du maximum de vraisemblance, nous égalons aux valeurs observées correspondantes les quantités suivantes : 1) dénombrements ajustés d'ilot (au moyen de l'effet principal pour l'ilot, i), 2) taux de réponse par ilot (au moyen du terme $r * t$), 3) caractéristiques moyennes des ménages pour l'ensemble des ménages (pour les caractéristiques x_1

interactions marginales à l'interaction donnée sont inclus dans le modèle, de sorte que celui-ci contient les effets principaux pour l'expression du modèle x_1 , l'indicateur de réponse r , les indices d'îlot i et les interactions $i * x_2, i * r, r * x_3$ et $r * a * x_4$.

Puisque, dans (1), x_4 interagit avec le secteur, c'est-à-dire le niveau le plus faible d'agrégation pour lequel on dispose de données sur les non-répondants, cette expression du modèle devrait représenter une classification assez grossière des ménages n'incluant que les caractéristiques exactes au niveau du secteur. L'expression x_3 peut inclure des ménages les plus importants pour faire des imputations exactes au niveau du secteur. L'expression x_4 peut inclure un niveau plus élevé d'agrégation géographique pour lequel même, l'expression x_1 pourrait inclure le plus grand nombre d'interactions, y compris l'interaction de toutes les variables qui définissent le type de ménage, puisqu'elle est ajustée au niveau d'agrégation géographique le plus élevé, qui utilise toutes les données disponibles. Enfin, x_2 , qui peut différer de x_3 , puisqu'elle interagit avec i au lieu de r , devrait être moins détaillée que x_1 , puisqu'elle interagit avec l'îlot, c'est-à-dire un niveau d'agrégation géographique beaucoup plus faible. Ces lignes directrices sont motivées par le fait que les estimations des interactions avec i , r ou a sont déterminées d'après un nombre relativement faible d'observations et devrait demeurer simple. Le choix de x_2, x_3 et x_4 comme il est décrit plus haut devrait améliorer la précision des estimations par modèle, tout en maintenant les taux de marge les plus importants.

Pour illustrer ce que pourraient être les termes x_1, \dots, x_4 les totaux de marge les plus importants.

Alors, une spécification possible de x_1, x_2 et x_3 est $x_1 = \text{race} * \text{taille} * \text{mode d'occupation}, x_2 = \text{race} * \text{taille} + \text{mode d'occupation}, x_3 = \text{taille} * \text{mode d'occupation}, \text{ et } x_4 = \text{race} + \text{taille} + \text{mode d'occupation}$. Permettre que les termes x_1, \dots, x_4 soient des expressions du modèle, plutôt que de simples interactions nous donne un moyen concis de représenter un modèle contenant toutes les interactions souhaitées. Par exemple, un modèle contenant un terme $i * x_2$, où x_2 correspond à la spécification susmentionnée, comprend à la fois une interaction îlot \times mode d'occupation.

Une interprétation heuristique de notre modèle logarithmique est que nous estimons la distribution détaillée des types de ménage sur la région entière (x_1), puis que nous déplaçons cette distribution pour tenir compte des caractéristiques générales de l'îlot (x_2), des différences générales entre les ménages répondants et non-répondants (x_3), ainsi que des différences les plus importantes entre les ménages répondants et non-répondants dans le secteur étudié (x_4).

Le modèle logarithmique contient des facteurs géographiques emboîtés pour les îlots et les secteurs. Il contient aussi des facteurs croisés représentant les caractéristiques démographiques des ménages, à savoir un indicateur de réponse à la première étape (ménage répondant ou non-répondant), un indice de type de ménage et des expressions du modèle en les variables qui définissent les types de ménage. Ces expressions du modèle sont des sous-modèles contenant toutes les interactions qui définissent le type de ménage (c'est-à-dire $\text{race} \times \text{taille} \times \text{mode d'occupation}$).

Nous utilisons la notation suivante :

i = indice d'îlot ($i = 1, \dots$, nombre d'îlots dans le DR);

j = indice de type de ménage ($j = 1, \dots$, nombre de types);

$a = a(i)$ = indice indiquant le secteur qui contient les îlots ($a = 1, \dots$, nombre de secteurs);

$x_k = x_k(j)$ = expressions du modèle en les variables qui définissent les types de ménage où $k = 1, 2, 3, 4$

x_1, \dots, x_4 est une expression du modèle qui est marginale à x_3 . (Cette terminologie est expliquée plus bas).

Nous supposons que le modèle logarithmique a la forme suivante :

$$n_{ijr}^{np} \sim \text{Poisson}(m_{ijr}^{np}), \log(m_{ijr}^{np}) = z_{ijr}^T \beta \quad (1)$$

où n_{ijr}^{np} et m_{ijr}^{np} sont, respectivement, les dénombrements observés et prévus pour l'îlot i , le type de ménage j et la situation de réponse r , et Z est la matrice de plan d'expérience correspondant à la formule du modèle suivante:

$$x_1 + i * x_2 + i * r + r * x_3 + r * a * x_4. \quad (2)$$

Conformément à la notation type de Wilkinson et Rogers (1973) pour les modèles linéaires généralisés, l'opérateur « * » indique que tous les effets principaux et toutes les

non-répondants non échantillonnés de chaque type au moyen d'une combinaison de modèles logistiques et loglinéaires. Cette étape est celle qui est l'objet du présent article (et de celui de FIT).

Pour la modélisation, nous avons classé les ménages par types en nous basant sur quelques caractéristiques importantes. Ici, nous utilisons 19 types, dont l'un est « logement inoccupé ». Les 18 autres sont définis par classification croisée des ménages pour trois catégories de taille (1 à 2 personnes, 3 à 4 personnes, 5 personnes et plus), trois catégories de race (Hispaniques, non-Hispaniques noirs et Autre) et deux catégories de mode d'occupation du logement (propriétaire, locataire).

Pour prédire le nombre de logements inoccupés parmi les logements non-répondants non échantillonnés dans chaque îlot, nous (et FIT) avons ajusté un modèle de régression logistique, en tenant compte du fait que la relation entre les ménages répondants et non-répondants n'est pas la même pour les logements inoccupés que pour ceux qui sont occupés. Les logements inoccupés répondants sont simplement ceux qui ont été considérés comme étant inoccupés par un facteur des services postaux, ce qui a entraîné le retour du questionnaire original. Leur distribution dépend vraisemblablement en grande partie des caractéristiques du logement associées à la distribution du courrier, ce qui nous renseigne peu sur la distribution des logements inoccupés non-répondants.

Après avoir modélisé les logements inoccupés, nous avons ajusté un modèle loglinéaire afin de prédire la distribution des types de ménage des logements occupés parmi les ménages non-répondants non échantillonnés restants à trois niveaux de détail géographique. L'îlot est la plus petite unité pour laquelle les dénombrements sont estimés. Le « domaine d'estimation » est l'unité la plus grande, c'est-à-dire la région dans laquelle l'estimation est faite indépendamment d'autres domaines de ce genre; dans notre application aux données de Recensement de 1990, il s'agit de la région pour laquelle le recensement a été réalisé à partir de l'un des 449 bureaux locaux de district, ou district de recensement (DR), comptant environ 200 000 ménages, en moyenne. Enfin, nous appelons « secteur » un niveau d'agrégation géographique intermédiaire comprenant un ensemble relativement homogène d'îlots contigus à l'intérieur d'un domaine d'estimation. Dans le contexte de la classification géographique type du Census Bureau, il pourrait s'agir de secteurs de recensement, de groupes d'îlots ou de secteurs du registre des adresses.

Nous exposons brièvement les dernières étapes qui seraient suivies pour obtenir les produits du recensement au moyen des estimations. À la deuxième étape de la méthode d'imputation, les dénombrements prévus seraient arrondis en nombres entiers. Des mécanismes sans biais (c'est-à-dire

3.2. Modèle loglinéaire

Nous avons ajusté un modèle loglinéaire pour estimer la prévalence des divers types de ménage parmi les ménages non-répondants non échantillonnés dans un DR, en utilisant des données provenant des répondants ainsi que des non-répondants dans l'échantillon SCNR pour ce DR. Le modèle prédit, pour chaque îlot, les types de ménage parmi les ménages non-répondants non échantillonnés, d'après des renseignements sur les caractéristiques des ménages répondants dans le même îlot et les caractéristiques des ménages non-répondants, déterminées d'après l'échantillon SCNR, dans les îlots voisins. À cette fin, le modèle loglinéaire contient des interactions entre les caractéristiques des ménages qui définissent le type de ménage et la situation de réponse à divers niveaux de détail géographique.

Cette stratégie de modélisation est motivée par le fait que, quand un modèle loglinéaire hiérarchique (c'est-à-dire un modèle dans lequel, pour chaque effet d'interaction inclus, les effets ou interactions principaux qui lui sont inclus, les valeurs ajustées pour le maximum de vraisemblance, les valeurs ajustées pour chaque total de marge ou moyenne correspondant à un effet dans le modèle sont égales aux totaux de marge ou aux moyennes observées correspondants (Birch 1963). Par conséquent, les prédictions pour les types de ménage concordent avec les taux observés pour les caractéristiques incluses dans le modèle, aux niveaux de détail géographique et situations de réponse correspondant aux interactions incluses dans le modèle. En outre, comme les prédictions du modèle pour les effets inclus sont contraintes de concorder avec les

des procédures stochastiques qui, en prédiction, imputent le nombre prédit d'unités dans chaque cellule, d'« arrondissement contrôlé » (c'est-à-dire arrondissement dans un tableau à double entrée en préservant les totaux de marge) ont été établis par Cox (1987), ainsi que par George et Penny (1987). Cependant, d'autres études doivent être réalisées pour déterminer s'il est possible de modifier ces méthodes de sorte que les nombres de ménages soient arrondis en assurant le maintien de toutes les valeurs de marge correctes-pendant aux effets inclus dans le modèle loglinéaire. Il s'agit d'un domaine où la recherche est active étant donné son importance en ce qui concerne la non-divulgaration statistique.

Ensuite, les renseignements détaillés sur les individus et les ménages seraient imputés pour les ménages non-répondants en leur substituant des ménages donneurs ayant les mêmes caractéristiques. Les donneurs peuvent être choisis parmi les non-répondants échantillonnés, les répondants, ou une combinaison des deux sources. Enfin, des totalisations et des échantillons de microdonnées seraient préparés à partir des listes complètes.

modèles de régression de Poisson (Bell et Otto 1994), ou les modèles loglinéaires plus complexes (tels que ceux proposés ici et dans Zanutto et Zaslavsky 1995b) a) sont utilisés pour estimer les dénombrements pour de petits domaines et pour des groupes démographiques de faible niveau d'agrégation pour lesquels des estimations directes sont impossibles. Comme nous, FIT classifient les ménages en un nombre modéré de types définis d'après des caractéristiques importantes (par exemple, nombre de personnes, race, mode d'occupation du logement), puis estiment le nombre de ménages de chaque type parmi les non-répondants non échantillonnés. Ils produisent ensuite une liste de recensement complète en imputant le nombre estimé de ménages de chaque type. La différence principale entre notre approche et celle de FIT tient au fait que l'utilisation d'un modèle loglinéaire plutôt qu'un modèle de ratio stratifié nous donne plus de souplesses en ce qui concerne la finesse des contraintes imposées à divers niveaux de détail géographique. Bell et Otto (1994) estiment le nombre de personnes de 18 ans et plus de chaque race (Hispaniques, non-Hispaniques, Noirs, Autre) dans chaque logement non-répondant non échantillonné, mais n'examinent pas la façon dont il faut grouper les personnes imputées dans les ménages ni la façon d'imputer les caractéristiques du ménage, comme le mode d'occupation du logement. Ces modèles « descendants » *ad hoc* ne contiennent, au plus, que quelques caractéristiques des ménages et, par conséquent, ne modélisent pas explicitement leur structure, mais ils sont conçus de façon à assurer la cohérence des agrégats jugés les plus importants.

Scharfer (1995) élabore une stratégie « ascendante » dans laquelle les ménages sont constitués en partant des individus, de leurs caractéristiques et de leurs relations, qui doivent toutes être exprimées par un modèle particulier. Ces modèles décrivent la population de façon plus détaillée et permettent de faire des inférences probabilistes (c'est-à-dire bayésiennes) complètes au sujet des caractéristiques observées. Cependant, contrairement à l'autre, cette approche oblige à construire un ensemble assez complexe de modèles avant que toute imputation puisse être faite. De surcroît, dans ce cadre, il est plus difficile d'assurer la convergence entre les microdonnées et les contrôles agrégés. Cependant, une stratégie combinée permettrait d'utiliser nos modèles pour produire des estimations quasiment sans biais selon le type de ménage et ceux de Scharfer pour exécuter les imputations.

3. Méthodes d'estimation et modèles

3.1 Vue d'ensemble

À la première étape de la méthode d'imputation, nous prédisons pour chaque îlot les nombres de ménages

(plus nombreuses) aux niveaux plus agrégés de détail géographique. Pour cela, nous commençons par classer les ménages en un petit nombre de types. Puis, nous utilisons un modèle loglinéaire hiérarchique pour estimer dans chaque îlot la distribution des types de ménage parmi les ménages non-répondants non échantillonnés. Cette distribution dépend des caractéristiques des ménages répondants du même îlot qui ont répondu par la poste et des ménages non-répondants échantillonnés dans les îlots voisins. Nous pouvons alors imputer les ménages non-répondants non échantillonnés d'après cette distribution estimée des types de ménage.

Bien que l'échantillonnage SCNR n'ait pas été utilisé lors du Recensement de 2000 pour des raisons juridiques complexes, notre stratégie d'estimation et d'imputation peut être adoptée pour l'estimation pour petits domaines ou l'imputation lors de tout recensement ou enquête avec échantillonnage SCNR où les unités sont mises en grappes de telle sorte que les caractéristiques des non-répondants soient reliées à celles des répondants compris dans le même domaine, ainsi qu'à celles des non-répondants échantillonnés dans les domaines voisins. Les méthodes apparemment de Purcell et Kish (1980) et de Zhang et Chambers (2004) reposent aussi sur l'utilisation de modèles loglinéaires pour estimer des dénombrements croisés sur petits domaines en supposant que les totaux de population sont connus et que du petit domaine. Nous possédons une source supplémentaire d'information, c'est-à-dire les caractéristiques des ménages non-répondants dans l'échantillon SCNR. Ceci nous permet de modéliser directement la relation entre les ménages répondants et non-répondants dans certains îlots.

À la section 2, nous résumons les stratégies proposées pour imputer les données manquantes dans cette situation. À la section 3.1, nous décrivons notre méthode générale d'échantillonnage et d'estimation. À la section 3.2, nous présentons notre modèle d'estimation et d'imputation et à la section 3.3, nos méthodes de lissage et d'estimation. À la section 4, nous évaluons notre modèle par simulation. Enfin, à la section 5, nous résumons les méthodes d'estimation de l'EQM et à la section 6, nous présentons nos conclusions.

2. Propositions antérieures pour l'imputation des non-répondants au recensement

Plusieurs méthodes ont été proposées pour imputer les caractéristiques des logements non-répondants. Les approches « descendantes » consistent à estimer d'abord des dénombrements pour des agrégats de ménages, puis à les répartir entre les petits domaines en maintenant la convergence avec les agrégats. Les modèles de ratio simple (Fuller, Isaki et Tsay 1994, nommés ci-après « FIT »), les

Un modèle d'estimation et d'imputation des ménages du recensement non-répondant sous échantillonnage pour le suivi des cas de non-réponse

Elaine L. Zanutto et Alan M. Zaslavsky¹

Résumé

L'échantillonnage pour le suivi des cas de non-réponse (échantillonnage SCNR) est une innovation qui a été envisagée lors de l'élaboration de la méthodologie du recensement décennal des États-Unis de 2000. L'échantillonnage SCNR consiste à envoyer des recenseurs auprès d'un échantillon seulement des ménages qui n'ont pas répondu au questionnaire initial envoyé par la poste; ce qui réduit les coûts, mais crée un problème important d'estimation pour petits domaines. Nous proposons un modèle permettant d'imputer les caractéristiques des ménages qui n'ont pas répondu au questionnaire envoyé par la poste, afin de profiter des économies importantes que permet de réaliser l'échantillonnage SCNR, tout en obtenant un niveau de précision acceptable pour les petits domaines. Notre stratégie consiste à modéliser les caractéristiques des ménages en utilisant un petit nombre de covariables aux niveaux agrégés de détail géographique et des covariables plus détaillées (plus nombreuses) aux niveaux plus agrégés de détail géographique. Pour cela, nous commençons par classer les ménages en un petit nombre de types de ménage parmi les ménages non-répondants qui ont retourné le questionnaire par la poste appartenant au même lot et des caractéristiques des ménages répondants qui ont retourné le questionnaire par la poste appartenant au même lot et des échantillons d'après cette distribution estimée des types de ménage. Nous évaluons les propriétés de notre modèle logarithme par simulation. Les résultats montrent que, comparativement aux estimations produites par des modèles de rééchantillonnage, notre modèle logarithme produit des estimations dont l'EQM est nettement plus faible dans de nombreux cas et à peu près la même dans la plupart des autres cas. Bien que l'échantillonnage SCNR n'ait pas été utilisé lors du recensement de 2000, notre stratégie d'estimation et d'imputation peut être appliquée lors de tout recensement ou enquête recourant cet échantillonnage dans les secteurs voisins.

Mots clés : Données manquantes; estimation pour petits domaines; ajustement proportionnel itératif; modèle logarithmique; ECM.

1. Introduction

L'échantillonnage pour le suivi des cas de non-réponse (SCNR) est une innovation qui a été envisagée lors de l'élaboration de la méthodologie du recensement décennal des États-Unis de 2000 (U.S. Bureau of the Census 1997a, b). Selon les procédures suivies à l'heure actuelle pour 99 % des ménages, le Censur Bureau commence par envoyer par la poste un questionnaire qui doit être retourné par la poste dûment rempli. Puis, des recenseurs essaient de prendre contact avec tous les ménages qui n'ont pas répondu par la poste (environ 35 % de ceux auxquels le questionnaire a été envoyé par la poste). La charge de travail que représente la communication avec ces quelques 42 millions de ménages fait de cette opération de suivi l'une des plus coûteuses du recensement.

L'échantillonnage SCNR comporte l'envoi de recenseurs auprès d'un échantillon seulement de ménages non-répondants. Cet échantillon est un échantillon non mis en grappes de logements non-répondants (l'« unité d'échantillonnage ») ou un échantillon en grappes comprenant toutes les unités

petites zones correspondant à un lot urbain ou à une région rurale compacte comptant environ 15 logements). Cette deuxième étape de suivi se solde par l'obtention d'un questionnaire rempli (par procuration ou imputation, au besoin) pour tous les logements échantillonnés, sauf ceux pour lesquels il est déterminé qu'ils sont inoccupés.

L'économie que permet de réaliser l'échantillonnage est importante, mais l'approche nécessite l'estimation des caractéristiques d'un très grand nombre de ménages non-répondants non échantillonnés, ce qui pose un problème considérable d'estimation pour petits domaines (Ghosh et Rao 1994; Rao 2003). Nous montrons qu'en utilisant des modèles appropriés pour imputer les caractéristiques des ménages non-répondants non échantillonnés, nous pouvons profiter des économies importantes de l'échantillonnage SCNR, tout en obtenant un niveau de précision acceptable pour les estimations sur petits domaines. Notre stratégie consiste à modéliser les caractéristiques des ménages en utilisant un petit nombre de covariables aux niveaux élevés de détail géographique et des covariables plus détaillées

1. Elaine L. Zanutto, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, États-Unis. Courriel : zanutto@wharton.upenn.edu; Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, États-Unis. Courriel : zaslavsky@hcp.med.harvard.edu

- Brick, J.M., et Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.
- Deville, J.C., et Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.
- Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, 429-440.
- Haziza, D., et Rao, J.N.K. (2005). Inference for domains under imputation for missing survey data. *Canadian Journal of Statistics*, 33, 149-161.
- Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 2, 169-174.
- Rao, J.N.K. (1990). Variance estimation under imputation for missing data. Rapport technique, Statistique Canada, Ottawa.
- Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of American Statistical Association*, 91, 499-506.
- Rao, J.N.K. (2005). Evaluation de l'interaction entre la théorie et la pratique des enquêtes par sondage. *Techniques d'enquête*, 31, 127-151.
- Rao, J.N.K., et Shao, J. (1992). On variance estimation under imputation for missing data. *Biometrika*, 79, 811-822.
- Rao, J.N.K., et Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
- Särndal, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.
- Shao, J., et Steel, P. (1999). Variance Estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.
- Sitter, R., et Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *Canadian Journal of Statistics*, 25, 61-73.

$$\hat{\mathbf{Y}}_{1(d)} = \hat{\mathbf{T}}^{-1} \left[\sum_{i=1}^s w_i a_i \mathbf{x}_i \mathbf{y}_i + \sum_{i=1}^s w_i (1 - a_i) \mathbf{x}_i \mathbf{y}_i^* \right]$$

$$= \hat{\mathbf{T}}^{-1} \sum_{i=1}^s w_i a_i \mathbf{x}_i \mathbf{y}_i^* \quad (41)$$

où $\hat{\mathbf{T}} = \sum_{i=1}^s w_i \mathbf{x}_i \mathbf{x}_i^t$. Notons que l'estimateur imputé $\hat{\mathbf{Y}}_{1(d)}$

dans (41) ne nécessite pas les identificateurs de réponse, a_i .

Haziza et Rao (2005) ont montré que l'estimateur imputé $\hat{\mathbf{Y}}_{1(d)}$ est biaisé sous l'hypothèse MN. Ils ont proposé un

estimateur corrigé pour le biais qui est approximativement

sans biais sous l'hypothèse MN ou sous l'hypothèse MI.

Nous proposons ici une extension de l'estimateur corrigé

pour le biais de Haziza-Rao qui est approximativement sans

biases sous l'hypothèse MNG ou sous l'hypothèse MI.

Il est facile de voir que, sous l'hypothèse MNG, le biais

de non-réponse conditionnel de l'estimateur imputé (41)

basé sur l'imputation par la régression déterministe modifiée

(18) est de la forme

$$\text{Biais}(\hat{\mathbf{Y}}_{1(d)} | s) \approx -\hat{\mathbf{T}}^{-1} \left[\sum_{i=1}^s w_i (1 - p_i) \mathbf{x}_i (\mathbf{y}_i - \mathbf{z}_i^t \mathbf{y}_{i,N}^*) \right] \quad (42)$$

où $\mathbf{y}_{i,N}^*$ est donné par (15). Un estimateur

approximativement conditionnellement sans biais du biais

exprimé par (42) est de la forme

$$B(\hat{\mathbf{Y}}_{1(d)} | s) \approx -\hat{\mathbf{T}}^{-1} \left[\sum_{i=1}^s \tilde{w}_i a_i \mathbf{x}_i (\mathbf{y}_i - \mathbf{z}_i^t \mathbf{y}_i^*) \right] \quad (43)$$

où \mathbf{y}_i^* est donné par (17). Un estimateur corrigé pour le

biais, $\hat{\mathbf{Y}}_{1(d)}^*$, est alors obtenu sous la forme $\hat{\mathbf{Y}}_{1(d)}^* -$

$B(\hat{\mathbf{Y}}_{1(d)} | s)$, qui mène à

$$\hat{\mathbf{Y}}_{1(d)}^* = \hat{\mathbf{T}}^{-1} \left[\sum_{i=1}^s \frac{\tilde{d}_i}{w_i} a_i \mathbf{x}_i (\mathbf{y}_i - \mathbf{z}_i^t \mathbf{y}_i^*) + \sum_{i=1}^s w_i \mathbf{x}_i \mathbf{z}_i^t \mathbf{y}_i^* \right] \quad (44)$$

L'estimateur corrigé pour le biais (44) est

approximativement sans biais sous l'hypothèse MI ou sous

l'hypothèse MNG. Donc, il est robuste au sens de sa validité

sous ces deux hypothèses. Cependant, il nécessite les

identificateurs de réponse a_i ainsi que les probabilités de

réponse \tilde{d}_i , contrairement à l'estimateur imputé $\hat{\mathbf{Y}}_{1(d)}$ dans

taux global de réponse. Dans ce cas, l'estimateur corrigé

$$\hat{\mathbf{Y}}_{1(d)}^* = \hat{\mathbf{T}}^{-1} \sum_{i=1}^s w_i \mathbf{x}_i \mathbf{z}_i^t \mathbf{y}_i^* + (1 - \hat{\mathbf{T}}^{-1}) \sum_{i=1}^s w_i \mathbf{x}_i \mathbf{z}_i^t \mathbf{y}_i^* \quad (45)$$

en notant que $\hat{\mathbf{y}}_i^* = \mathbf{y}_i^*$, où, sous l'imputation par la

régression déterministe,

$$\hat{\mathbf{y}}_i^* = \left(\sum_{i=1}^s w_i \mathbf{z}_i \mathbf{z}_i^t / (\lambda_i \mathbf{z}_i) \right)^{-1} \times \left[\sum_{i=1}^s w_i a_i \mathbf{z}_i \mathbf{y}_i / (\lambda_i \mathbf{z}_i) + \sum_{i=1}^s w_i (1 - a_i) \mathbf{z}_i \mathbf{y}_i^* / (\lambda_i \mathbf{z}_i) \right]$$

$$= \mathbf{y}_i^* \quad (45)$$

Haziza et Rao (2005) ont obtenu l'estimateur corrigé pour le

biais (45).

Conclusion

Par souci de simplicité, nous avons considéré le cas

d'une classe d'imputation unique, mais notre méthode

MNG s'étend facilement à plusieurs classes d'imputation au

moyen d'imputations distinctes dans le cas de plusieurs

classes. Par exemple, nous pourrions procéder à l'impu-

tation par la moyenne pondérée dans les classes en utilisant

nos poids modifiés \tilde{w}_i . En outre, notre méthode peut être

étendue au cas de l'imputation composite (Sitter et Rao

1997; Shao et Steel 1999) qui repose sur diverses impu-

tations pour les réponses manquantes à une question selon

le modèle du ratio reliant \mathbf{y} à \mathbf{x} ne sera pas applicable,

contrairement au cas où la variable \mathbf{x} est observée sur toutes

les unités échantillonnées.

Remerciements

Les travaux de recherche de J.N.K. Rao ont été financés

par une bourse du Conseil de recherches en sciences

naturelles et en génie du Canada. Les auteurs remercient les

examineurs de leurs commentaires et suggestions utiles.

Bibliographie

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a

quasi-model-assisted approach. *Journal of the Royal Statistical*

Society, B, 67, 445-458.

Binder, D.A. (1983). On the variances of asymptotically normal

estimators from complex surveys. *Revue Internationale de*

Statistique, 15, 279-292.

Statistique Canada, N° 12-001-XPB au catalogue

inférieur à -10 %. Donc, l'estimateur de la variance le plus simple $v_{naïve}$ pourrait convenir en pratique.

Tableau 5

Biais relatif (%) des estimateurs

de la variance

f	$BR(v_{naïve})$	$BR(v_{correct})$
0,05	-6,3	-5,1
0,10	-5,8	-4,1
0,25	-4,3	-3,2

5. Estimation de moyennes de domaine

En pratique, il est souvent nécessaire de produire des

estimations pour divers domaines (sous-populations). Par exemple, dans le cas de l'enquête sur la population active du Canada, des estimations du chômage sont requises selon le groupe âge-sexe et selon l'industrie au niveau provincial. Pour corriger pour la non-réponse partielle, on pourrait utiliser la méthode d'imputation par la régression modifiée proposée. Cependant, les domaines doivent être spécifiés d'avance à l'étape de l'imputation. Autrement dit, les indicateurs de domaine doivent faire partie du modèle d'imputation. Or, en pratique, les domaines ne sont généralement pas spécifiés à l'étape de la vérification et de l'imputation, si bien que les estimations par domaine sont calculées d'après des données imputées fondées sur des modèles d'imputation ne contenant pas les indicateurs de domaine. Par conséquent, les estimateurs imputés utilisés pour les domaines sont généralement biaisés. Nous proposons un estimateur corrigé pour le biais, s'inspirant de la section 2.2, pour remédier à ce problème. L'estimateur corrigé pour le biais peut être obtenu à l'étape de l'estimation et ne nécessite pas la spécification des domaines à l'étape de l'imputation.

Nous pouvons exprimer un vecteur de moyennes de domaine sous la forme

$$\underline{Y}^{(p)} = \left(\sum_{i=1}^U \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^U \mathbf{x}_i y_i, \quad (39)$$

où $\mathbf{x} = (x_{1i}, \dots, x_{di}, \dots, x_{Di})'$ est un vecteur d'indicateurs de domaine, $x_{di} = 1$ si $i \in$ domaine d et $x_{di} = 0$, autrement. Nous supposons que \mathbf{x} est connu pour toutes les unités $i \in s$. Autrement dit, seule la réponse à la variable y peut être manquante. En l'absence de non-réponses, un estimateur approximativement sans biais de $\underline{Y}^{(p)}$ est donné par

$$\hat{\underline{Y}}^{(p)} = \left(\sum_{i=1}^s w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^s w_i \mathbf{x}_i y_i. \quad (40)$$

En présence de non-réponse à la question y , un estimateur imputé de $\underline{Y}^{(p)}$ est donné par

l'imputation aléatoire serait estimée au moyen de (37) avec p_i remplacé par \hat{p}_i .

4.3. Étude par simulation

Nous avons réalisé une petite étude par simulation afin d'évaluer les propriétés des estimateurs de la variance

engendrés aux sections 4.1 et 4.2. Nous avons généré une population de taille $N = 2\,500$ contenant deux variables y et z . Nous avons d'abord généré la variable z à partir d'une loi Gamma avec un paramètre d'échelle égal à 4 et un paramètre de forme égal à 10. Puis, nous avons produit les valeurs de y conformément au modèle du ratio

$$y_i = \gamma z_i + e_i,$$

où les e_i sont générées à partir d'une loi normale de moyenne 0 et de variance σ^2 . Nous avons fixé la valeur du paramètre γ à 2 et avons choisi la variance σ^2 de façon que le R^2 du modèle soit environ égal à 0,81. L'objectif est d'estimer le total de population $Y = \sum y_i$.

Nous avons généré $R = 10\,000$ échantillons aléatoires simples sans remise à partir de la population finie en utilisant les fractions d'échantillonnage n/N suivantes : 0,05, 0,1 et 0,25. Dans chaque échantillon, nous avons généré la non-réponse à la question y selon le mécanisme de réponse suivant : la probabilité de réponse p_i pour l'unité i est donnée par le modèle logistique

$$\log \frac{p_i}{1 - p_i} = \lambda_0 + \lambda_1 z_i.$$

Nous avons choisi les valeurs de λ_0 et λ_1 de façon à

obtenir un taux de réponse global d'environ 70 %. Nous avons généré les indicateurs de réponse a_i indépendamment à partir d'une loi de Bernoulli de paramètre p_i .

Pour corriger pour la non-réponse à la variable y , nous avons utilisé l'imputation par le ratio déterministe modifiée pour laquelle les valeurs imputées sont données par (19). D'après chaque échantillon simulé, nous avons calculé l'estimateur imputé \hat{Y}_i donné par (2) avec les valeurs imputées (19). Comme mesure du biais de l'estimateur de la variance v , nous avons utilisé le biais relatif $[E(v) - EQM(\hat{Y}_i)]/EQM(\hat{Y}_i)$. Nous désignons par $v_{naïve}$ l'estimateur de la variance totale obtenu par sommation de (34) et (36) quand les probabilités de réponse p_i sont remplacées par les probabilités de réponse estimées \hat{p}_i , et par $v_{correct}$ l'estimateur de la variance totale obtenu par sommation de (36) et (36) avec p_i remplacé par \hat{p}_i . Le tableau

5 donne le biais relatif (en %) des deux estimateurs de la variance. Il montre clairement que ces estimateurs donnent lieu à une sous-estimation, mais que celle-ci est un peu moins prononcée dans le cas de $v_{correct}$. En outre, ils donnent tous deux de bons résultats, le biais relatif étant

recensement correspondant à \mathbf{u} et γ_i , respectivement. Un estimateur de θ donné par $\hat{\theta} = (\hat{\eta}_i, \hat{\gamma}_i', \hat{\gamma}_i')$ peut être exprimé comme une solution des équations d'estimations au

niveau de l'échantillon

$$\hat{S}(\theta) = 0,$$

$$\text{où } \hat{S}(\theta) = (\hat{S}_1(\theta), \hat{S}_2(\theta), \hat{S}_3(\theta))' \text{ avec}$$

$$\hat{S}_1(\theta) = \sum_{i=1}^s w_i' \mathbf{u}_i' \mathbf{u}_i' (a_i' - f(\mathbf{u}_i' \mathbf{u}_i' N)) = \mathbf{0},$$

$$\hat{S}_2(\theta) = \sum_{i=1}^s w_i' a_i' z_i' \frac{f(\mathbf{u}_i' \mathbf{u}_i' N)}{(1 - f(\mathbf{u}_i' \mathbf{u}_i' N))} (Y_i - \mathbf{z}_i' \gamma_i' N) / (\lambda_i' z_i) = \mathbf{0}$$

et

$$\hat{S}_3(\theta) = Y - \sum_{i=1}^s w_i' \gamma_i' N - \sum_{i=1}^s w_i' a_i' (Y_i - \mathbf{z}_i' \gamma_i' N) = 0.$$

Soit $\hat{\mathbf{J}}(\theta) = (\partial \hat{S}(\theta) / \partial \theta)$ la matrice de dimensions $(k + l + 1) \times (k + l + 1)$ des dérivées partielles. Nous avons

$$\mathbf{V}(\theta) = [\hat{\mathbf{J}}^{-1}(\theta)] \hat{\Sigma}(\theta) [\hat{\mathbf{J}}^{-1}(\theta)]',$$

où $\hat{\Sigma}(\theta)$ dénote la matrice symétrique de dimensions $(k + l + 1) \times (k + l + 1)$ dont l'élément $\hat{\sigma}_{ij}$ est la covariance entre $\hat{S}_i(\theta)$ et $\hat{S}_j(\theta)$ en ce qui a trait à l'échantillonnage, sachant le vecteur des indicateurs de réponse \mathbf{a} . Si $\hat{\Sigma}(\theta)$ est remplacé par un estimateur convergent $\hat{\Sigma}(\theta)$, disons, nous obtenons un estimateur de la variance convergent $\mathbf{V}(\theta)$ donné par

$$\mathbf{V}(\theta) = [\hat{\mathbf{J}}^{-1}(\theta)] \hat{\Sigma}(\theta) [\hat{\mathbf{J}}^{-1}(\theta)]'.$$

Puisque nous nous intéressons à l'estimateur de la variance, γ_i , de \hat{Y}_i , nous avons besoin de la ligne finale, \mathbf{b} , disons, de $\hat{\mathbf{J}}^{-1}(\theta)$, évaluée à $\theta = \theta$. Il s'ensuit que

$$\gamma_i = \mathbf{b}' \hat{\Sigma}^{-1}(\theta) \mathbf{b}, \quad (38)$$

Pour obtenir la composante γ_i , nous supposons que les poids d'échantillonnage w_i satisfont $\max(w_i / N w_i) = O(1)$ et qu'il existe une constante positive C telle que $C < p_i$. En outre, nous supposons que $\hat{\eta} - \eta = O_p(n^{-1/2})$. Par linéarisation de Taylor, nous obtenons

$$\hat{Y}_i = \hat{Y}_{ip} + (\eta - \eta) \sum_{i=1}^s p_i^{-1} (Y_i - \hat{Y}_{ip}) + O_p(N/n),$$

où

$$\hat{Y}_{ip} = \left[\sum_{i=1}^U (1 - a_i) \mathbf{z}_i' \mathbf{z}_i' / (\lambda_i' \mathbf{z}_i) \right] \left[\sum_{i=1}^U (1 - a_i) \mathbf{z}_i' \gamma_i / (\lambda_i' \mathbf{z}_i) \right].$$

En supposant que $f(\mathbf{u}_i' \mathbf{u}_i) / \partial \eta$ est bornée uniformément, nous avons

$$E_p(\hat{Y}_i) = E_p(\hat{Y}_{ip}) + O_p(N/n^{1/2}).$$

Donc, la composante $E_p[E_p(\hat{Y}_{ip}) - Y | \mathbf{a}]$ est donnée approximativement par (35) et γ_i est donné par (36) avec p_i remplacé par \hat{p}_i . Dans le cas de l'imputation par la régression aléatoire modifiée, la composante due à

Nous obtenons alors la composante γ_i par estimation des quantités inconnues dans (35), ce qui mène à

$$\gamma_i = \sum_{i=1}^s w_i' a_i' (1 - p_i') \hat{\zeta}_i^2, \quad (39)$$

où

$$\hat{\zeta}_i^2 = \left[1 + \frac{p_i}{(1 - p_i')} \frac{1}{\lambda_i' \mathbf{z}_i'} (\mathbf{Z} - \mathbf{Z}_i) \hat{\mathbf{T}}^{-1} \mathbf{z}_i' \right] (Y_i - \mathbf{z}_i' \gamma_i' N).$$

Un estimateur de la variance totale γ_i est obtenu par sommation de (34) et de (36) : $\gamma_i = \gamma_i + \gamma_i$. En pratique, les probabilités de réponse sont inconnues. Par conséquent, il est impossible de calculer l'estimateur de la variance γ_i .

Une solution simple à remplacer p_i par les probabilités de réponses estimées \hat{p}_i dans (34) et (36), puis à utiliser l'estimateur résultant γ_i comme estimateur de la variance de \hat{Y}_i . Comme nous le montrons à l'aide d'une étude en simulation à la section 4.3, cette méthode simple donne des résultats acceptables.

4.1.2 Imputation par la régression aléatoire modifiée

Nous commençons par noter que

$$Y_i (Y_i^*) =$$

$$(\lambda_i' \mathbf{z}_i') \sum_{i=1}^s w_i' \frac{(1 - p_i')}{(1 - p_i')} a_i' (e_j - e_j^2) \left/ \sum_{i=1}^s w_i' \frac{p_i'}{(1 - p_i')} a_i' \right. \equiv s_i^2$$

et $\text{Cov}(Y_i^*, Y_j^*) = 0, i \neq j$. Donc, d'après (2), la composante γ_i , due à l'imputation aléatoire, est donnée par

$$\gamma_i = \sum_{i=1}^s w_i' (1 - a_i) Y_i^* = \sum_{i=1}^s w_i' (1 - a_i) s_i^2. \quad (37)$$

Un estimateur de la variance totale est obtenu par sommation de (34), (36) et (37) : $\gamma_i = \gamma_i + \gamma_i + \gamma_i$. De nouveau, puisque les probabilités de réponse p_i sont inconnues, il est impossible de calculer γ_i dans (37). Nous proposons de remplacer dans cette équation les p_i par les probabilités de réponse estimées \hat{p}_i .

4.2 p_i inconnues

Nous utilisons la méthode de Binder (1983) pour dériver la composante γ_i lorsque les probabilités de réponse

est un vecteur de dimension l de paramètres auxiliaires disponibles pour tout $i \in s$. Par exemple, dans le cas de la régression logistique, $f(\mathbf{u}_i' \mathbf{u}_i) = \exp(\mathbf{u}_i' \mathbf{u}_i) / \exp(1 + \mathbf{u}_i' \mathbf{u}_i)$. Les probabilités de réponse estimées sont données par $\hat{p}_i = f(\mathbf{u}_i' \mathbf{u}_i)$, où $\hat{\eta}$ est un estimateur convergent de η . Soit $\theta = (\eta_1^N, \gamma_1^N, \gamma_1^N)$, où η_1^N et γ_1^N sont les paramètres de

Tableau 4
Biais relatif (%) et RREQM (%) des estimateurs imputés

Estimateur imputé*	Nombre de classes	BR	RREQM
\hat{y}_f^p (mécanisme 1)	1	14,4	14,5
	5	-0,02	4,26
	10	-0,85	7,33
	20	-0,20	8,61
	30	-0,03	8,61
	40	0,03	9,09
	50	0,06	9,44
\hat{y}_f^p (mécanisme 1)	—	1,11	1,90
\hat{y}_f^p (mécanisme 2)	1	29,0	29,1
	5	21,4	21,4
	10	21,0	21,1
	20	20,9	21,0
	30	20,9	21,0
	40	21,0	21,0
	50	21,0	21,0
\hat{y}_f^p (mécanisme 2)	—	10,9	10,9

* \hat{y}_f^p est donné par (27) et \hat{y}_f^p est donné par (28).

4. Estimation de la variance

À la présente section, nous établissons un estimateur de la variance de l'estimateur imputé \hat{Y}_f , en suivant l'approche renversée de Fay (1991). La variance totale de \hat{Y}_f sous une méthode d'imputation déterministe particulière est donnée par

$$V(\hat{Y}_f - Y) = E'V'E^p(\hat{Y}_f - Y | \mathbf{a}), \quad (29)$$

où $\mathbf{a} = (a_1, \dots, a_N)'$ est le vecteur des indicateurs de réponse (Shao et Steel 1999). Un estimateur de la variance globale $V(\hat{Y}_f - Y)$ dans (29) est donné par $v_f = v_1 + v_2$, où

$$v_1 \text{ est un estimateur de } V^p(\hat{Y}_f - Y | \mathbf{a}) \text{ étant donné les indicateurs de réponse } a_i, \text{ et } v_2 \text{ est un estimateur de } V[E^p(\hat{Y}_f - Y | \mathbf{a})], \text{ l'estimateur } v_1 \text{ ne dépend pas du mécanisme de réponse ni du modèle d'imputation et, par conséquent, } v_1 \text{ est valide sous l'hypothèse MNG ou sous l'hypothèse ML.}$$

Sous l'imputation aléatoire correspondante, la variance de l'estimateur imputé \hat{Y}_f est donnée par

$$V(\hat{Y}_f - Y) = E'V'E^pE_*(\hat{Y}_f - Y | \mathbf{a}) + V'E^pE_*(\hat{Y}_f - Y | \mathbf{a}), \quad (30)$$

où $V_*(\cdot)$ représente l'opérateur de variance en ce qui a trait à l'imputation aléatoire. Nous supposons que $E_*(\hat{Y}_f | \mathbf{a})$ coïncide avec l'estimateur imputé pour le cas déterministe. Donc, $E'V'E^pE_*(\hat{Y}_f - Y | \mathbf{a})$ est estimé par v_1 dans ce cas. La contribution supplémentaire à la variance due à l'imputation aléatoire vient de la composante $V'E^pE_*(\hat{Y}_f - Y | \mathbf{a})$ qui est estimée par v_2 . Par conséquent, il découle de (30) que la variance globale $V(\hat{Y}_f - Y)$ est estimée par $v_f = v_1 + v_2$. Le terme v_2 est absent dans le cas de l'imputation déterministe.

4.1 p_i connues

4.1.1 Imputation par la régression déterministe modifiée

Sous l'imputation par la régression déterministe modifiée, l'estimateur imputé quand les p_i sont connues peut s'écrire

$$\hat{Y}_{fp} = \sum_{i=1}^s w_i a_i y_i + (\hat{Z} - \hat{Z}_r)' \hat{Y}_{rp}, \quad (31)$$

$$\hat{Y}_{rp} = \left[\sum_{i=1}^s w_i a_i \frac{(1-p_i)^d}{(1-p_i)^d} \mathbf{z}_i' \mathbf{z}_i / (\lambda_i' \mathbf{z}_i) \right]^{-1} \left[\sum_{i=1}^s w_i a_i \frac{d_i}{(1-p_i)^d} \mathbf{z}_i' \mathbf{z}_i / (\lambda_i' \mathbf{z}_i) \right]. \quad (32)$$

Pour obtenir v_1 , nous appliquons la méthode de linéarisation de Taylor standard qui donne

$$\hat{Y}_{fp} - Y \approx \sum_{i=1}^s w_i \tilde{\tilde{z}}_{ip}, \quad (33)$$

$$\tilde{\tilde{z}}_{ip} = a_i y_i + (1 - a_i) \mathbf{z}_i' \hat{Y}_{rp} + (\hat{Z} - \hat{Z}_r)' \mathbf{T}_{i-1}^d a_i \frac{d_i}{(1-p_i)^d} \mathbf{z}_i' \mathbf{z}_i / (\lambda_i' \mathbf{z}_i) + \sum_{i=1}^s w_i a_i \frac{d_i}{(1-p_i)^d} \mathbf{z}_i' \mathbf{z}_i / (\lambda_i' \mathbf{z}_i).$$

Si nous dénotons l'estimateur de la variance de l'estimateur en échantillon complet $\hat{Y} = \sum_{i=1}^s w_i y_i$ par $v(Y)$, il découle de (33) qu'un estimateur de $V^p(\hat{Y}_f - Y | \mathbf{a})$ est donné par

$$v_1 = v(\tilde{\tilde{z}}_p), \quad (34)$$

que nous obtenons en remplaçant y_i par $\tilde{\tilde{z}}_{ip}$ dans la formule de $v(Y)$.

Pour obtenir la deuxième composante v_2 , commençons par noter que

$$E^p(\hat{Y}_{fp} - Y | \mathbf{a}) \approx \sum_{i=1}^s a_i y_i + \sum_{i=1}^N (1 - a_i) Y_p - Y,$$

$$Y_p = \left[\sum_{i=1}^N a_i \frac{d_i}{(1-p_i)^d} \mathbf{z}_i' \mathbf{z}_i / (\lambda_i' \mathbf{z}_i) \right]^{-1} \left[\sum_{i=1}^N a_i \frac{d_i}{(1-p_i)^d} \mathbf{z}_i' \mathbf{z}_i / (\lambda_i' \mathbf{z}_i) \right].$$

En appliquant la linéarisation de Taylor, nous pouvons montrer que

$$V^p[E^p(\hat{Y}_{fp} - Y | \mathbf{a})] \approx \sum_{i=1}^N p_i (1 - p_i) \tilde{\tilde{z}}_{ip}^2, \quad (35)$$

pour la seconde, on emploie uniquement les variables auxiliaires reliées avec la variable d'intérêt.

Tableau 3

Biais relatif (%) et RREQM (%) des estimateurs imputés sous le mécanisme de réponse 2

Scénario	Biais (classique) (proposé)	RREQM (classique) (proposé)
$Y(1) - P(1)$	1,84	2,55
$Y(2) - P(1)$	4,46	1,84
$Y(3) - P(1)$	2,03	2,70
$Y(4) - P(1)$	-4,58	5,07
$Y(1) - P(2)$	1,84	2,55
$Y(4) - P(2)$	-1,70	5,07
$Y(1) - P(3)$	1,84	2,55

3.2 Étude par simulation 2

Nous avons généré une population finie de taille $N = 1\,000$ contenant trois variables : une variable d'intérêt y et trois variables auxiliaires z_1, z_2 et z_3 , en commençant par générer z_1, z_2 et z_3 indépendamment à partir d'une loi exponentielle de moyenne 100, puis en générant les valeurs de y selon le modèle de régression

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i} + \epsilon_i,$$

où les ϵ_i sont générées à partir d'une loi normale de moyenne 0 et de variance σ^2 . Les valeurs des paramètres $\gamma_0, \gamma_1, \gamma_2$ et γ_3 ont été fixées, respectivement, à 20, 10, 0,5 et 10. Nous avons choisi la variance σ^2 de façon que le R^2 du modèle soit approximativement égal à 0,66. L'objectif est d'estimer la moyenne de population $\bar{Y} = \sum U y_i / N$. Afin de nous concentrer sur l'erreur due à la non-réponse/imputation, nous avons considéré le cas d'un recensement, c'est-à-dire $n = N = 1\,000$. Ensuite, nous avons généré la non-réponse à la variable y pour la population simulée selon les mécanismes de réponse suivants :

Mécanisme 1 : La probabilité de réponse p_{1i} de l'unité i est donnée par le modèle logistique

$$\log \frac{p_{1i}}{1 - p_{1i}} = \lambda_0 + \lambda_1 z_{1i} + \lambda_2 z_{3i}.$$

Mécanisme 2 : La probabilité de réponse p_{2i} de l'unité i est donnée par le modèle logistique

$$\log \frac{p_{2i}}{1 - p_{2i}} = \lambda_0 + \lambda_1 y_i + \lambda_2 z_{3i}.$$

Nous avons choisi les valeurs de λ_0, λ_1 et λ_2 de façon à obtenir un taux de réponse global d'environ 70 %. Les indicateurs de réponse a_{1i} et a_{2i} ont été générés indépendamment $R = 1\,000$ fois à partir d'une loi de Bernoulli avec les paramètres p_{1i} et p_{2i} , respectivement.

Nous avons utilisé deux stratégies pour corriger pour la non-réponse. La première consiste à diviser l'échantillon, s , en classes d'imputation s_1, s_2, \dots, s_C en nous fondant sur les variables auxiliaires z_1, z_2 et z_3 . Pour former les classes, nous avons utilisé la méthode du score qui peut se décrire comme il suit. En utilisant l'information auxiliaire, nous avons d'abord estimé les probabilités de réponse, p_i , pour obtenir p_i pour les répondants ainsi que les non-répondants par régression logistique sur z_1, z_2 et z_3 . Puis, en utilisant les p_i , nous avons divisé la population en C classes en suivant la procédure FASTCLUS de SAS (qui utilise pour la classification l'algorithme des k moyennes). La méthode du score mène à une partition de la population telle que, dans les classes, les unités (répondants et non-répondants) sont homogènes par rapport aux valeurs p_i . La deuxième stratégie s'appuie sur la méthode d'imputation par la régression modifiée proposée basée sur les variables auxiliaires z_1, z_2 et z_3 . Le but de l'étude en simulation est de comparer les propriétés des deux estimateurs imputés de la moyenne de population \bar{Y} : a) Estimateur imputé fondé sur les C classes d'imputation :

$$\bar{y}_I^c = \sum_{c=1}^C \frac{N_c}{N} \bar{y}_{I^c}, \quad (27)$$

où

$$\bar{y}_{I^c} = \frac{1}{N_c} \left[\sum_{i=1}^{s_c} w_i a_i y_i + \sum_{i=1}^{s_c} w_i (1 - a_i) y_i^* \right],$$

et $N_c = \sum_{i=1}^{s_c} w_i$. Nous avons utilisé l'imputation par la moyenne pondérée dans les classes, c'est-à-dire $y_i^i = \sum_{i=1}^{s_c} w_i a_i y_i / \sum_{i=1}^{s_c} w_i a_i$.

b) Estimateur imputé fondé sur l'imputation par la régression modifiée proposée, dénoté \bar{y}_I :

$$\bar{y}_I = \frac{1}{N} \left[\sum_{i=1}^s w_i a_i y_i + \sum_{i=1}^s w_i (1 - a_i) y_i^* \right], \quad (28)$$

Souignons que, dans cette étude en simulation, $w_i = 1$ pour tout $i \in U$, parce qu'aucun échantillonnage n'a eu lieu. Enfin, le tableau 4 donne une comparaison de ces estimateurs en ce qui concerne le biais relatif, donné par (23), et de la RREQM, donnée par (25). L'examen du tableau 4 montre clairement que l'estimateur imputé proposé (28) donne de nettement meilleurs résultats que l'estimateur (27) fondé sur les classes d'imputation en ce qui a trait à la RREQM, pour le mécanisme 1 ainsi que le mécanisme 2.

Afin de compenser pour la non-réponse à la variable y , nous avons utilisé l'imputation par la régression déterministe classique pour laquelle les valeurs imputées sont données par (6) et l'imputation par la régression déterministe modifiée pour laquelle les valeurs imputées sont données par (18). Les imputations ont été fondées sur les modèles de y et de p énumérés au tableau 1, c'est-à-dire $y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)}$ et $p^{(1)}, p^{(2)}, p^{(3)}$. Notons que $p^{(1)}$ correspond au mécanisme de réponse 1 et $y^{(1)}$, au modèle générant la population.

Tableau 1

Modèles utilisés pour l'imputation			
Modèles pour y	Ordonnée à l'origine	z_1	z_2
$y^{(1)}$	Oui	Oui	Oui
$y^{(2)}$	Oui	Non	Oui
$y^{(3)}$	Oui	Non	Oui
$y^{(4)}$	Ordonnée à l'origine	z_1	z_2
Modèles pour p_i			
$p^{(1)}$	Oui	Oui	Oui
$p^{(2)}$	Oui	Non	Oui
$p^{(3)}$	Non	Oui	Non

D'après chaque échantillon simulé, nous avons calculé l'estimateur imputé \hat{y}_i donné par (2) avec les valeurs imputées (6) et (18), en nous basant sur certaines combinaisons des modèles $y^{(a)}$ et $p^{(a)}$, $a = 1, \dots, 4$. $b = 1, 2, 3$. Comme mesure du biais d'un estimateur imputé \hat{y}_i , nous avons utilisé le biais relatif (BR) simulé exprimé en pourcentage donné par

$$BR(\hat{y}_i) = \frac{\text{Biais}(\hat{y}_i)}{Y} \times 100, \quad (23)$$

$$\text{Biais}(\hat{y}_i) = \frac{1}{R} \sum_{r=1}^R \hat{y}_i^{(r)} - Y \quad (24)$$

$$RREQM(\hat{y}_i) = \sqrt{\frac{\text{Biais}(\hat{y}_i)}{\text{Biais}(\hat{y}_i)}} \times 100, \quad (25)$$

$$EQM(\hat{y}_i) = \frac{1}{R} \sum_{r=1}^R (\hat{y}_i^{(r)} - Y)^2. \quad (26)$$

Les résultats concernant le biais relatif et la RREQM sont présentés au tableau 2 pour les échantillons générés selon le mécanisme de réponse 1 et au tableau 3 pour ceux générés

selon le mécanisme de réponse 2. L'examen du tableau 2 montre clairement que, si l'imputation est effectuée confortablement au modèle correct (c'est-à-dire $y^{(1)}$), l'imputation par la régression déterministe classique mène à un estimateur approximativement sans biais et est plus efficace que l'imputation par la régression déterministe modifiée en ce qui concerne la RREQM. Comme l'a souligné un examinateur, l'imputation par la régression déterministe modifiée peut produire des estimateurs plus efficaces que la régression déterministe classique. Autrement dit, il existe des scénarios (non examinés ici) pour lesquels la méthode d'imputation par la régression déterministe modifiée proposée pourrait être plus efficace que la méthode d'imputation par la régression déterministe classique. Quand le modèle d'imputation est spécifié incorrectement (c'est-à-dire $y^{(2)}$ et $y^{(4)}$), l'imputation déterministe produit des estimateurs avec biais, tandis que l'imputation déterministe modifiée induit un biais faible à négligeable, à condition que le modèle de non-réponse soit spécifié correctement (c'est-à-dire $p^{(1)}$). Par conséquent, la RREQM est plus grande dans le cas de l'imputation déterministe classique que dans celui de l'imputation déterministe modifiée. Si les modèles d'imputation et de non-réponse sont tous deux spécifiés incorrectement (c'est-à-dire $y^{(4)} - p^{(2)}$), tous les estimateurs sont biaisés.

Tableau 2

Biais relatif (%) et RREQM (%) des estimateurs imputés sous le mécanisme de réponse 1

Scénario	Biais (classique)	Biais (proposé)	RREQM (classique)	RREQM (proposé)
$y^{(1)} - p^{(1)}$	0,19	-0,01	1,85	2,33
$y^{(2)} - p^{(1)}$	5,20	0,16	5,60	2,66
$y^{(3)} - p^{(1)}$	0,17	-0,04	1,87	2,37
$y^{(4)} - p^{(1)}$	-14,80	-3,50	15,00	6,70
$y^{(1)} - p^{(2)}$	0,19	0,12	1,85	1,86
$y^{(4)} - p^{(2)}$	-14,80	-14,80	15,00	14,60
$y^{(1)} - p^{(3)}$	0,19	0,05	1,85	1,88

Si l'on examine le tableau 3, il est évident que, dans le cas du mécanisme 2, l'estimateur imputé obtenu sous imputation par la régression modifiée donne des résultats aussi bons, voire meilleurs, que l'estimateur imputé obtenu sous imputation par la régression classique dans tous les scénarios. Ce résultat n'est pas étonnant, puisque, pour arriver à réduire efficacement le biais dans le cas de la non-réponse non ignorable, il est nécessaire d'utiliser toute l'information auxiliaire appropriée disponible. Or, l'information auxiliaire utilisée dans le cas de l'imputation par la régression modifiée est plus riche que celle utilisée pour l'imputation proposée par la régression classique, puisque, pour la première, on se sert des variables auxiliaires reliées à la variable d'intérêt y ainsi que celles reliées à la probabilité de réponse, tandis que

2.4 Imputation par la régression aléatoire

L'imputation aléatoire peut être considérée comme une imputation déterministe avec ajout d'un bruit aléatoire. Soit s_m et s_m les ensembles de non-répondants et de répondants dans l'échantillon, respectivement, et soit $e_j = (y_j - z_j) / (\lambda_j^{1/2})$ les résidus standardisés pour les répondants $j \in s_r$ sous l'imputation par la régression déterministe. En outre, $e'_j = e_j$ avec $P(e'_j = e_j) = w_j / \sum w_i$ indépendamment pour chaque $i \in s_m$. Alors, l'imputation par la régression aléatoire utilise les valeurs imputées $y'_j = z_j + e'_j$, $i \in s_m$, où $e'_j = (\lambda_j z_j)^{1/2} (e'_j - e_j)$ avec $e'_j = \sum w_i e'_j / \sum w_i$. Soit $E^*(\cdot)$ l'espérance sous le processus d'imputation aléatoire. Nous avons $E^*(e'_j) = 0$ et $E^*(I_j^2)$ égale à (8). Par conséquent, l'estimateur imputé I_j^* est approximativement sans biais sous l'hypothèse MN ou sous l'hypothèse MII. Il convient de souligner que l'imputation par la régression aléatoire couvre le cas particulier de l'imputation hot-deck (pondérée) aléatoire. Pour le montrer, considérons le modèle d'imputation par la moyenne $E^*(w_i y_i) = \gamma$, $\text{Cov}^*(w_i y_i, y_j) = 0$, $i \neq j$. Nous avons $y'_j = \sum w_i a_i y_i / \sum w_i a_i = \bar{y}_j$, la moyenne pondérée des valeurs de y fournies par les répondants et $e'_j = y_j - \bar{y}_j$. Par conséquent, $y'_j = y_j + e'_j = y_j$ correspond à la valeur de y_j pour les répondants tirés aléatoirement avec probabilité

3. Etudes par simulation

Nous avons effectué deux études par simulation afin d'étudier les propriétés en échantillon fin des méthodes d'imputation par la régression déterministe modifiée et par la régression aléatoire modifiée proposées en termes de biais

L'estimateur imputé fondé sur l'imputation par la régression aléatoire est asymptotiquement entaché d'un biais sous l'hypothèse MNG. Pour obtenir un estimateur approximativement sans biais pour X , nous proposons une méthode d'imputation par la régression aléatoire modifiée. Soit $\hat{\sigma}_j = (Y_j - Z_j' \gamma_j) / (\lambda_j Z_j' Z_j)^{1/2}$ et $\hat{\sigma}_j^* = \hat{\sigma}_j$ avec $P(\hat{\sigma}_j^* = \hat{\sigma}_j) = w_j / \sum w_j a_j$, indépendamment pour chaque $i \in S_m$, où γ_j^* est donné par (17) et $w_j = w_j(1 - p_j) / p_j$. Alors, l'imputation par la régression aléatoire modifiée utilise les valeurs imputées $y_j^* = Z_j' \gamma_j^* + \hat{\sigma}_j^*$, où $\hat{\sigma}_j^* = (\lambda_j Z_j' Z_j)^{1/2} (\hat{\sigma}_j^* - \hat{\sigma}_j)$ avec $\hat{\sigma}_j^* = \sum w_j a_j \hat{\sigma}_j / \sum w_j a_j$. Nous avons $E_j(\hat{\sigma}_j^*) = 0$ et $E_j(\gamma_j^*)$ égale à l'estimateur imputé sous l'imputation par la régression déterministe modifiée. Donc, l'estimateur imputé \hat{Y}_j est approximativement sans biais sous l'hypothèse MNG ou sous l'hypothèse ML. Pour les cas particuliers du modèle d'imputation par la moyenne, nous avons $\gamma_j^* = \sum w_j a_j \gamma_j / \sum w_j a_j$ et $y_j^* = y_j$ correspond à la valeur de y_j pour les répondants tirés aléatoirement avec probabilité $w_j / \sum w_j a_j$.

et de la racine relative de l'erreur quadratique moyenne. La première étude par simulation consiste à comparer les propriétés de l'imputation par la régression déterministe classique et de l'imputation par la régression déterministe modifiée proposée lorsque le modèle d'imputation et/ou le modèle de non-réponse ne sont pas spécifiés correctement. La deuxième a pour but de comparer les propriétés de l'estimateur imputé obtenues en utilisant des classes d'imputation fondées sur les probabilités de réponses estimées et de l'imputation par la moyenne pondérée (classique) à celles de l'estimateur imputé obtenu en utilisant la méthode d'imputation par la régression déterministe modifiée proposée.

3.1 Étude par simulation 1

Nous avons généré une population finie de taille $N = 1\,000$ contenant 3 variables : une variable d'intérêt y et deux variables auxiliaires z_1 et z_2 . Pour cela, nous avons commencé par générer z_1 et z_2 indépendamment à partir de lois exponentielles de moyenne 4 et 30, respectivement. Puis, nous avons généré les valeurs de y conformément au modèle de régression

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \epsilon_i$$

où les c_i sont générées à partir d'une loi normale de moyenne 0 et de variance σ^2 . Les valeurs des paramètres γ_0, γ_1 et γ_2 ont été fixées, respectivement, à 20, 2 et 0,1, et la variance σ^2 a été choisie de façon que le R^2 du modèle soit approximativement égal à 0,75. L'objectif est d'estimer le total de population $Y = \sum U Y_i$. Nous avons généré $R = 5\,000$ échantillons aléatoires simples sans remise de taille $n = 100$ à partir de la population finie. Dans chaque échantillon, la non-réponse à la variable y a été générée selon les mécanismes de réponse suivants :

Mécanisme 1 : La probabilité de réponse p_{1i} de l'unité i est donnée par le modèle de régression logistique

$${}^1\chi + {}^0\chi = \frac{{}^1d-1}{{}^1d} \text{sol}$$

Mécanisme 2 : La probabilité de réponse p_{2i} de l'unité i est donnée par le modèle de régression logistique

$$\log \frac{1 - d^{2l}}{d^{2l}} = \chi_0 + \chi_1 \chi_1^*.$$

Nous avons choisi les valeurs de λ_0 et λ_1 de façon que le taux de réponses global, soit approximativement de 70 %. Les indicateurs de réponse a_{1i} et a_{2i} ont été générés indépendamment à partir d'une loi de Bernoulli avec les paramètres p_{1i} et p_{2i} , respectivement. Notons que, dans le cas du mécanisme de réponse 2, le mécanisme de réponse est non ignorable en ce sens que la probabilité de réponse dépend de la variable d'intérêt.

Notons que $\tilde{y}_{s,N}$ est inconnu, puisque les valeurs de y ne sont observées que pour $i \in s$, et que les probabilités de réponses p_i sont inconnues. Un estimateur de $\tilde{y}_{s,N}$ fondé sur les unités répondantes et les probabilités de réponse estimées \hat{p}_i est donné par

$$\hat{y}_r = \left[\sum_{i \in s} w_i a_i \frac{\hat{p}_i}{(1 - \hat{p}_i)} z_i z_i' / (\lambda_i z_i) \right]^{-1} \sum_{i \in s} w_i a_i \frac{\hat{p}_i}{(1 - \hat{p}_i)} z_i y_i' / (\lambda_i z_i). \quad (17)$$

Nous avons $E(\hat{y}_r | s) \approx \tilde{y}_{s,N}$, de sorte que \hat{y}_r est conditionnellement apparemment sans biais pour $\tilde{y}_{s,N}$. Donc, en utilisant les valeurs imputées

$$y_i^* = z_i' \hat{y}_r \quad (18)$$

dans (2) avec \hat{y}_r , nous obtenons un estimateur apparemment sans biais du total X sous l'hypo-

thèse MNG. Notons que \hat{y}_r est un estimateur par les moindres carrés pondérés de y par rapport à un nouvel ensemble de poids, $\tilde{w}_i / (\lambda_i' z_i)$, où $\tilde{w}_i = w_i (1 - \hat{p}_i) / \hat{p}_i$. Donc, la procédure accroit les poids de sondage w_i des unités pour lesquelles $\hat{p}_i < 1/2$ et diminue ceux des unités pour lesquelles $\hat{p}_i > 1/2$. L'estimateur imputé peut être appliqué au fichier de données imputé contenant les poids d'échantillonnage w_i et les \hat{y}_i uniquement; les indicateurs de réponse a_i et les probabilités de réponse estimées \hat{p}_i ne sont pas requis. Toutefois, il est nécessaire de connaître a_i et \hat{p}_i pour estimer la variance. Notons que le producteur du fichier de données imputé utilise l'information concernant a_i et \mathbf{u}_i pour ajuster le modèle de réponse (4) et générer les valeurs imputées y_i^* données par (18).

L'utilisation des valeurs imputées (18) mène également à un estimateur approximativement sans biais de X sous l'hypothèse MI. Premièrement, sous le modèle de régression (3), en notant que $E_m(y_i | s) = z_i' \gamma$ et $E_m(\hat{y}_i | s) = \gamma$, nous avons $E_m(\hat{y}_i - y_i | s) = 0$ et $E_r E_m(\hat{y}_i - y_i | s) = 0$ sans spécifier le mécanisme de non-réponse MAR sous-jacent. Donc, l'utilisation des valeurs imputées (18) mène à un estimateur robuste au sens de sa validité sous les deux approches. Enfin, il est intéressant de souligner que les valeurs imputées (18) peuvent également être obtenues par la méthode d'imputation par calage (Beaumont 2005). Cette dernière consiste à trouver des valeurs imputées finales aussi proche que possible des valeurs imputées originales conformément à une fonction de distance, sous les contraintes de calage.

Deux cas particuliers de l'imputation par la régression modifiée (18) présentent un intérêt, à savoir i) l'imputation par la moyenne modifiée avec $z_i = z_i'$ et $\lambda_i' z_i = z_i'$ et ii) l'imputation par la moyenne modifiée avec $z_i = z_i'$ et $\lambda_i' z_i = 1$ et qui est un cas particulier de l'imputation par le ratio.

Supposons que les poids d'échantillonnage w_i satisfont $\max(n/N w_i) = O(1)$ et qu'il existe une constante positive C telle que $C < p_i$. Alors,

$$\hat{y}_r^{\text{opt}} = \sum_{i \in s} w_i' (1 - \hat{p}_i) z_i' / \left(\frac{1}{n} \right) + O \left(\frac{1}{n} \right).$$

Donc, pour les grandes tailles d'échantillon, le choix \hat{y}_r^{opt} est presque optimal pour l'imputation par le ratio. De même, $\hat{y}_{s,N}$ est presque optimal pour l'imputation par la moyenne, qui est un cas particulier de l'imputation par le ratio.

est l'estimateur par les moindres carrés pondérés de γ sous le modèle (3), basé sur les unités échantillonnées répondant à la question γ . Partant de (6), l'estimateur imputé (2) peut s'écrire sous la forme

$$\hat{\gamma} = \hat{\gamma}_r + (\hat{\mathbf{Z}} - \hat{\mathbf{Z}}_r)' \hat{\gamma}_p, \quad (8)$$

où $\hat{\gamma}_r = \sum_s w_r a_r \gamma_r$, $\hat{\mathbf{Z}} = \sum_s w_r \mathbf{z}_r$, et $\hat{\mathbf{Z}}_r = \sum_s w_r a_r \mathbf{z}_r$. Notons que l'estimateur imputé (8) est similaire à un estimateur par la régression dans le cas de l'échantillonnage à deux phases.

Sous l'hypothèse MN, $E_r(a_r | s) = p$ et le biais de non-réponse conditionnel, $E_r(\hat{\gamma}_r | s) - \gamma$, est approximativement égal à 0. En outre, sous l'hypothèse MI et le modèle de régression (3), le biais de non-réponse conditionnel $E_r E_m(\hat{\gamma}_r | s) - \gamma$, est nul. Cependant, sous l'hypothèse MNG, le biais de non-réponse conditionnel est donné par

$$E_r(\hat{\gamma}_r | s) - \gamma | s \approx - \sum_s w_r (1 - p_r) (\gamma_r - \mathbf{z}_r' \hat{\gamma}_p) = B(\hat{\gamma}_r | s), \quad (9)$$

où

$$\hat{\gamma}_p = \left(\sum_{i=1}^s w_i p_i \mathbf{z}_i \mathbf{z}_i' / (\lambda_i \mathbf{z}_i) \right)^{-1} \sum_{i=1}^s w_i p_i \mathbf{z}_i \gamma_i / (\lambda_i \mathbf{z}_i), \quad (10)$$

le terme de (9) disparaît, en notant que $(\sum_s w_i \mathbf{z}_i) \hat{\gamma}_p = \lambda' (\sum_s w_i \mathbf{z}_i \mathbf{z}_i' / (\lambda_i \mathbf{z}_i)) \hat{\gamma}_p = \lambda' (\sum_s w_i \mathbf{z}_i \gamma_i / (\lambda_i \mathbf{z}_i)) = \sum_s w_i \gamma_i$.

2.2 Estimateur corrigé pour le biais

Nous supposons pour le moment que les probabilités de réponse p_i sont connues. Une approche naturelle en vue d'éliminer le biais présent dans $\hat{\gamma}_r$ sous l'hypothèse MNG consiste à considérer un estimateur corrigé pour le biais de la forme

$$\hat{\gamma}_o^r = \hat{\gamma}_r - B(\hat{\gamma}_r | s), \quad (11)$$

où $B(\hat{\gamma}_r | s)$ est un estimateur de $B(\hat{\gamma}_r | s)$:

$$B(\hat{\gamma}_r | s) = - \sum_s w_r a_r (\hat{\gamma}_r | s) = - \sum_s w_r a_r (\hat{\gamma}_r | s) / (1 - p_r). \quad (12)$$

Soulignons que $E_r(\hat{\gamma}_r | s) \approx B(\hat{\gamma}_r | s)$ sous l'hypothèse MNG. En introduisant (12) dans (11) par substitution, nous obtenons un estimateur corrigé pour le biais de la forme

$$\hat{\gamma}_o^r = \sum_s w_r \frac{p_r}{1 - p_r} a_r \gamma_r + \left(\sum_s w_r \mathbf{z}_r' - \sum_s w_r \frac{p_r}{1 - p_r} a_r \mathbf{z}_r' \right) \hat{\gamma}_p. \quad (13)$$

Notons que (13) est également sous la forme d'un estimateur par la régression dans le cas de l'échantillonnage à deux phases.

En pratique, les probabilités de réponse p_i sont inconnues. Supposons que nous puissions obtenir des estimateurs \hat{p}_i de p_i par modélisation de p_i conformément au modèle de non-réponse (4). Alors, nous obtenons un

estimateur corrigé pour le biais en remplaçant p_i par \hat{p}_i dans (13). Cet estimateur est également approximativement conditionnellement sans biais sous l'hypothèse MI. Donc, l'estimateur corrigé pour le biais (13) est robuste au sens de sa validité sous l'hypothèse MNG. Cependant, contrairement à l'estimateur imputé $\hat{\gamma}_r$ donné par (2), l'estimateur corrigé pour le biais $\hat{\gamma}_o^r$ ne peut être calculé sans que l'on connaisse les identificateurs de réponse, a_i , et les probabilités de réponse estimées, \hat{p}_i . Par conséquent, pour pouvoir obtenir $\hat{\gamma}_o^r$, les indicateurs de réponse ainsi que les probabilités estimées de réponse doivent être fournis avec le fichier de données imputé, ce qui n'est pas toujours le cas en pratique. Cet inconvénient de $\hat{\gamma}_o^r$ peut être éliminé en utilisant la nouvelle méthode d'imputation, décrite à la section 2.3, qui mène à un estimateur approximativement sans biais sous l'hypothèse MNG ou sous l'hypothèse MI sans que l'on connaisse a_i et \hat{p}_i dans le fichier de données imputé. Cependant, il est nécessaire d'avoir accès aux valeurs de a_i et de \hat{p}_i pour estimer la variance.

2.3 Imputation par la régression déterministe modifiée

Nous supposons pour l'instant que les probabilités de réponse p_i sont connues. Nous utilisons alors les valeurs imputées

$$\gamma_i^* = \mathbf{z}_i' \hat{\gamma}_r^s \quad (14)$$

pour remplacer les valeurs manquantes γ_i et obtenons la forme de $\hat{\gamma}_s$ qui mène à un estimateur approximativement sans biais sous l'hypothèse MNG.

2.3.1 Estimateur approximativement sans biais

Le lemme qui suit donne la forme de $\hat{\gamma}_s$ qui mène à un estimateur approximativement sans biais sous l'hypothèse MNG.

Lemme 1 : Sous l'hypothèse MNG, le choix de $\hat{\gamma}_s$ qui mène à $E_r(\hat{\gamma}_r | s) - \gamma | s = 0$ est donné par

$$\hat{\gamma}_s^* = \left[\sum_{i=1}^s w_i (1 - p_i) \mathbf{z}_i \mathbf{z}_i' / (\lambda_i \mathbf{z}_i) \right]^{-1} \sum_{i=1}^s w_i (1 - p_i) \mathbf{z}_i \gamma_i / (\lambda_i \mathbf{z}_i). \quad (15)$$

Preuve : Le biais de non-réponse conditionnel de $\hat{\gamma}_r$ avec $\gamma_i^* = \mathbf{z}_i' \hat{\gamma}_r^s$ sous l'hypothèse MNG est donné par

$$E_r(\hat{\gamma}_r | s) - \gamma | s = - \sum_s w_r (1 - p_r) (\gamma_r - \mathbf{z}_r' \hat{\gamma}_r^s).$$

En notant que $(\lambda_r \mathbf{z}_r) / (\lambda_r \mathbf{z}_r) = 1$, il s'ensuit que $E_r(\hat{\gamma}_r | s) - \gamma | s = 0$ si $\hat{\gamma}_r^s$ satisfait

$$\left[\lambda_r' \sum_{i=1}^s w_i (1 - p_i) \mathbf{z}_i (\mathbf{z}_i' \gamma_i) / (\lambda_i \mathbf{z}_i) \right] = 0. \quad (16)$$

Le choix $\hat{\gamma}_r^s = \hat{\gamma}_s^*$ satisfait (16).

Nous supposons ici que la probabilité de réponse, p_i , de l'unité i est liée à un vecteur de dimension l de variables auxiliaires \mathbf{u}_i conformément à un modèle logistique de

$$p_i = f(\mathbf{u}_i' \boldsymbol{\eta}) = \exp(\mathbf{u}_i' \boldsymbol{\eta}) / \exp(1 + \mathbf{u}_i' \boldsymbol{\eta}), \quad (4)$$

où $\boldsymbol{\eta}$ est le vecteur de dimension l de paramètres du modèle. Le modèle (4) est le modèle de non-réponse hypothétique. Il peut être validé à partir des valeurs a_i et \mathbf{u}_i pour $i \in s$. Notons que a_i et \mathbf{u}_i sont particuliers à la variable d'intérêt. De plus, notons que l'hypothèse MN est un cas particulier de l'hypothèse MNG. Comme dans l'approche MN, des hypothèses explicites au sujet du mécanisme de réponse sont formulées et l'inférence sous l'hypothèse MNG est faite sous les conditions d'échantillonnage répété et du mécanisme de réponse hypothétique.

Rappelons que l'imputation est utilisée en vue de réduire le biais de non-réponse, en supposant que les variables auxiliaires disponibles permettent d'expliquer la variable pour laquelle des valeurs doivent être imputées et/ou la probabilité de réponse à la variable. Donc, en pratique, le choix de l'approche (MI ou MNG) devrait être dicté par la qualité du modèle d'imputation et du modèle de non-réponse. Le choix entre la modélisation de la probabilité de réponse à la variable et celle de la variable d'intérêt dépendra de la confiance que l'on a dans chacun des modèles. S'il peut paraître intuitivement plus séduisant de modéliser la variable d'intérêt, il existe, en pratique, des cas où il pourrait être plus facile de modéliser la probabilité de réponse (approche MNG). Par exemple, à Statistique Canada, l'Enquête sur les dépenses en immobilisations produit des données sur l'investissement fait au Canada, dans tous les types d'industries. Dans cette enquête, deux variables d'intérêt importantes sont les dépenses d'immobilisations en constructions neuves (CC) et les dépenses d'immobilisations en machines et matériel neufs (CM). Durant une année donnée, un grand nombre d'entreprises ne font aucun investissement en constructions neuves ni en machines neuves. Par conséquent, le fichier de données d'échantillon contient un grand nombre de valeurs nulles pour les variables CC et CM. Le cas échéant, la modélisation de la variable d'intérêt (CC ou CM) peut s'avérer difficile.

En général, les poids de sondage sont utilisés dans l'imputation par la régression linéaire. L'estimateur imputé ainsi obtenu d'un total de population est « robuste » au sens de l'absence de biais approximatif sous l'hypothèse MN ou sous l'hypothèse MI. Toutefois, il contient généralement un biais sous l'hypothèse MNG. Dans le présent article, nous proposons une nouvelle méthode d'imputation par la régression linéaire qui est robuste au sens où elle mène à des estimateurs approximativement sans biais sous l'hypothèse MNG ou sous l'hypothèse MI.

2. Estimation d'un total

Nous présentons aussi les résultats des simulations concernant les estimateurs de la variance. Enfin, à la section 5, nous examinons le cas des moyennes de domaine.

Population \rightarrow recensement avec non-répondants \rightarrow échantillon avec non-répondants.

À la section 2, nous décrivons l'élaboration d'une nouvelle méthode d'imputation par la régression linéaire déterministe, ainsi qu'un imputation par la régression linéaire aléatoire, et nous démontrons la propriété de robustesse dans le cas d'un total de population Y . À la section 3, nous présentons les résultats d'une étude par simulation des propriétés dans le cas d'échantillons finis de l'estimateur imputé sous la nouvelle méthode d'imputation. À la section 4, nous développons les estimateurs de la variance, en utilisant l'approche « renversé » de Fay (1991) dans laquelle l'ordre de l'échantillonnage et de la réponse est inversé :

À la présente section, nous étudions le biais de l'estimateur imputé Y_I . L'erreur totale, $Y_I - Y$, peut être décomposée comme il suit :

$$Y_I - Y = (Y_I - Y) + (Y_I - Y). \quad (5)$$

Dans (5), le terme $Y_I - Y$ est appelé erreur d'échantillonnage, tandis que le terme $Y_I - Y$ est appelé erreur due à la non-réponse/imputation. Soulignons qu'il n'y a pas d'erreur due à l'imputation dans le cas d'imputation déterministe. Puisque l'erreur d'échantillonnage ne dépend ni de la non-réponse ni de la méthode d'imputation, nous nous concentrons sur l'erreur due à la non-réponse/imputation $Y_I - Y$ et évaluons ses propriétés étant donné l'échantillon s . Sous l'approche MN ou MNG, nous définissons le biais de non-réponse conditionnel comme étant $E_p(Y_I - Y | s)$, où $E_p(\cdot)$ représente l'espérance par rapport au mécanisme de réponse. Sous l'approche MI, le biais de non-réponse conditionnel est défini comme étant $E_p(E_m(Y_I - Y | s))$ sous l'hypothèse MAR.

2.1 Imputation par la régression déterministe

L'imputation par la régression déterministe consiste à utiliser les valeurs imputées

$$y_i^* = \mathbf{z}_i' \boldsymbol{\gamma}^* \quad (6)$$

pour remplacer les valeurs manquantes y_i , où

$$\boldsymbol{\gamma}^* = \left(\sum_{i=1}^s w_i a_i \mathbf{z}_i \mathbf{z}_i' / (\lambda' \mathbf{z}_i) \right) \left(\sum_{i=1}^s w_i a_i \mathbf{z}_i y_i / (\lambda' \mathbf{z}_i) \right) \quad (7)$$

avec les poids d'échantillonnage (ou de sondage) $w_i = 1/\pi_i$, où π_i dénote la probabilité d'inclusion de l'unité i dans l'échantillon $s, i = 1, \dots, N$. Rao (2005) a suggéré que l'on devrait appeler (1) l'estimateur de Narain-Horvitz-Thompson (NHT) en reconnaissance du fait que Narain (1951) a également découvert (1) indépendamment d'Horvitz et de Thompson (1952).

En présence de non-réponse à la variable y_j , nous utilisons l'imputation et définissons un estimateur imputé \tilde{Y}_j de la

forme

$$\tilde{Y}_j = \sum_{s=1}^S w_s a_j y_{js} + \sum_{s=1}^S w_s (1 - a_j) y_{js}^* = \sum_{s=1}^S w_s \tilde{y}_{js}, \quad (2)$$

où y_{js}^* représente la valeur imputée pour remplacer la valeur manquante y_{js}, a_{js} représente l'indicateur de réponse égal à 1 si l'unité i répond à la variable y et égale à 0 autrement, et $\tilde{y}_{js} = a_{js} y_{js} + (1 - a_{js}) y_{js}^*$. L'estimateur imputé (2) peut être obtenu à partir du fichier de données imputé contenant les poids de sondage w_i et les \tilde{y}_{js} uniquement, sans que l'on connaisse les indicateurs de réponse a_{js} . Cependant, ces derniers seront nécessaires pour estimer la variance. Soit $p_i = P(a_i = 1)$ la probabilité de réponse de l'unité i à la variable y . Dans le présent article, nous supposons que les unités répondent indépendamment les unes des autres, c'est-à-dire que $p_{ij} = P(a_i = 1, a_j = 1) = p_i p_j$, si $i \neq j$.

Comme toute méthode de remplacement des données manquantes, l'imputation nécessite certaines hypothèses au sujet du mécanisme de réponse et/ou du modèle d'imputation. En présence de données imputées, deux approches sont généralement utilisées pour mener des inférences au sujet des totaux, des moyennes et d'autres paramètres d'intérêt, à savoir i) l'approche du modèle d'imputation (MI) et ii) celle du modèle de non-réponse (MNR). L'approche (i) est également appelée approche assistée par un modèle (Särndal 1992) et l'approche (ii), approche fondée sur le plan de sondage (Shao et Steel 1999). L'approche MN est basée sur le partitionnement de la population U en J classes d'imputation, suivi de l'imputation des valeurs de y correspondant aux non-répondants comptés dans chaque classe en utilisant les valeurs de y des répondants comptés dans la même classe comme donneurs, indépendamment dans chacune des J classes. L'hypothèse suivante est formulée :

Hypothèse MN : La probabilité de réponse à une variable d'intérêt est constante dans les classes d'imputation. Autrement dit, $p_i = p_j$, disons, où l'indice inférieur v désigne la classe d'imputation.

Dans l'approche MN, des hypothèses explicites sont formulées au sujet du mécanisme de réponse. Il s'ensuit que l'inférence sous l'hypothèse MN est faite sous les conditions d'échantillonnage répété et d'un mécanisme de réponse uniforme dans les classes. L'approche MN a été étudiée par Rao (1990, 1996), Rao et Shao (1992), Rao et

Sitter (1995) et Shao et Steel (1999), entre autres. Pour simplifier, nous supposons qu'il n'y a qu'une seule classe d'imputation, de sorte que $p_i = p$ sous l'hypothèse MN. L'approche MI est fondée sur l'hypothèse suivante :

Hypothèse MI : Les valeurs d'une variable manquent au hasard (MAR pour *missing at random*) au sens où la probabilité de réponse ne dépend pas de la valeur de la variable qui est imputée, mais des variables auxiliaires utilisées pour l'imputation. En outre, une hypothèse est émise quant au modèle qui génère les valeurs y_{ji} de la variable. Dans l'approche MI, des hypothèses explicites au sujet de la distribution des valeurs y_{ji} de la variable sont formulées au moyen d'un modèle appelé « modèle d'imputation ». Il s'ensuit que l'inférence sous l'hypothèse MI est faite sous les conditions d'échantillonnage répété et du modèle hypothétique qui génère la population finie de valeurs de y et de non-répondants à la variable y . Contrairement à l'approche MN, le mécanisme de réponse sous-jacent n'est pas spécifié, à part l'hypothèse MAR. L'hypothèse MI concernant le mécanisme de réponse est nettement plus faible que l'hypothèse MN de réponse uniforme dans les classes, mais les inférences sous l'hypothèse MI dépendent du modèle de population hypothétique. L'approche MI a été étudiée, entre autres, par Särndal (1992), Deville et Särndal (1994), ainsi que Shao et Steel (1999). Sous l'imputation par la régression linéaire, l'approche MI s'appuie sur le modèle d'imputation par la régression linéaire hypothétique suivant :

$$E_m(y_i) = \mathbf{z}_i' \boldsymbol{\gamma}, \quad V_m(y_i) = \sigma^2 = \sigma^2(\boldsymbol{\lambda}' \mathbf{z}_i), \\ \text{Cov}_m(y_i, y_j) = 0 \text{ si } i \neq j, \quad (3)$$

où $\boldsymbol{\gamma}$ est un vecteur de dimension k de paramètres inconnus, \mathbf{z}_i est un vecteur de dimension k de variables auxiliaires disponibles pour toutes $i \in s$, $\boldsymbol{\lambda}$ est un vecteur de dimension k de constantes spécifiques, σ^2 est un paramètre inconnu et E_m, V_m et Cov_m représentent, respectivement, les opérateurs d'espérance, de variance et de covariance par rapport au modèle d'imputation. La contrainte $\sigma^2 = \sigma^2(\boldsymbol{\lambda}' \mathbf{z}_i)$ ne restreint pas sévèrement la gamme de modèles d'imputation. Dans le présent article, nous proposons une troisième approche, appelée approche du modèle de non-réponse généralisée (MNG), qui est fondée sur l'hypothèse suivante :

Hypothèse MNG : Les valeurs de la variable manquent au hasard (MAR) et la probabilité de réponse est spécifiée sous forme d'une fonction des variables auxiliaires, \mathbf{u}_i , observées sur toutes les unités de l'échantillon, et de paramètres inconnus $\boldsymbol{\eta}$.

Une approche fondée sur un modèle de non-réponse à des fins d'inférence en présence d'imputation pour des données d'enquête manquantes

David Haziza et Jon N.K. Rao¹

Résumé

En présence de non-réponse partielle, deux approches utilisées à des fins d'inférence des paramètres d'intérêt. La première repose sur l'hypothèse que la réponse est ignorante dans les classes d'imputation, tandis que la seconde s'appuie sur l'hypothèse que la réponse est ignorable, mais utilise un modèle pour la variable d'intérêt comme mécanisme de réponse précise ignorable sans que doive être spécifié un modèle de la variable d'intérêt. Dans ce cas, nous montrons comment obtenir des valeurs imputées qui mènent à des estimateurs d'un total approximativement sans biais sous l'approche proposée, ainsi que sous la deuxième des approches susmentionnées. Nous obtenons aussi des estimateurs de la variance des estimateurs imputés qui sont approximativement sans biais en suivant une approche proposée par Fay (1991) dans laquelle sont inversés l'ordre de l'échantillonnage et de la réponse. Enfin, nous effectuons des études par simulation afin d'évaluer les propriétés des méthodes dans le cas d'échantillons finis, en termes de biais et d'erreur quadratique moyenne.

Mots clés : Approche basée sur un modèle de non-réponse; approche basée sur un modèle d'imputation; estimateur corrigé pour le biais; estimation de la variance; imputation par la régression aléatoire; imputation par la régression déterministe; non-réponse partielle.

1. Introduction

Il y a non-réponse partielle lors d'une enquête quand une unité échantillonnée participe à l'enquête, mais omet de répondre à une ou à plusieurs variables (Brick et Kalton 1996). Elle est généralement traitée par une forme ou l'autre d'imputation qui consiste à « boucher les trous » dues aux valeurs manquantes pour chaque variable. L'imputation peut effectivement réduire le biais, à condition que l'on dispose d'information auxiliaire appropriée pour toutes les unités échantillonnées et qu'on l'intègre correctement dans le modèle d'imputation et/ou dans le modèle de non-réponse.

L'imputation offre, entre autres, les caractéristiques suivantes : i) elle mène à la création d'un fichier de données complet et ii) elle permet d'utiliser les mêmes poids de sondage pour toutes les variables, ce qui assure que les résultats obtenus, après diverses analyses de l'ensemble complet de données, soient cohérents, contrairement aux résultats d'analyses réalisées sur un ensemble de données incomplet. Cependant, l'imputation présente aussi, entre autres, les difficultés suivantes : a) l'imputation marginale pour chaque variable fausse la relation entre les variables et b) traiter les valeurs imputées comme s'il s'agissait de valeurs réelles peut entraîner une sous-estimation im-portante de la variance des estimateurs imputés, partiellement quand le taux de non-réponse est appréciable. Des méthodes permettant de résoudre les problèmes (a) et (b) ont été proposées dans la littérature.

Dans le présent article, nous nous concentrons sur l'imputation marginale qui est utilisée communément dans de nombreuses enquêtes. Pour commencer, nous examinons de l'imputation par la régression linéaire déterministe qui comprend les cas particuliers de l'imputation par la moyenne et de l'imputation par le ratio. Selon cette méthode, une valeur manquante est remplacée par la valeur prédite obtenue en ajustant un modèle de régression linéaire au moyen des valeurs fournies par les répondants et de celles des variables auxiliaires recueillies pour toutes les unités échantillonnées. Nous examinons aussi le cas de l'imputation par la régression aléatoire qui peut être considérée comme une imputation par la régression déterministe à laquelle est ajouté un résidu aléatoire. Elle comprend le cas particulier de l'imputation hot-deck aléatoire. Soit U une population finie de taille éventuellement inconnue N . L'objectif est d'estimer le total de population $Y = \sum_{i \in U} y_i$ d'une variable y lorsque l'on a utilisé l'imputation pour traiter la non-réponse pour les valeurs y_i de la variable. Pour être concis, nous utiliserons la notation $\sum_{i \in A}$ pour $\sum_{i \in U}$, où $A \subseteq U$. Supposons que l'on sélectionne un échantillon probabiliste, s , de taille n conformément à un plan spécifié $p(s)$ à partir de U . Sous des conditions de réponse complète à la variable y , un estimateur de Y sans biais par rapport au plan est donné par l'estimateur d'Horvitz-Thompson bien connu

$$\hat{Y} = \sum_{i \in s} w_i y_i, \quad (1)$$

1. David Haziza, Division des méthodes d'enquête auprès des entreprises, Statistique Canada, Ottawa (Ontario) Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa (Ontario), Canada, K1S 5B6.

- Deville, J.-C., et Samdal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Dorfman, A.H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics*, 35, 29-41.
- Harms, T. (2003). Extensions of the calibration approach: calibration of distribution functions and its link to small area estimators. Chinese working paper #13, Federal Statistical Office, Allemagne.
- Kovachevic, M.S. (1997). Calibration estimation of cumulative of the Survey Methods Section, *Statistical Society of Canada*, 139-144.
- Kovar, J.G., Rao, J.N.K., et Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16 (Supp.), 25-45.
- Kuk, A.Y.C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika*, 75, 97-103.
- Kuk, A.Y.C., et Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Séries B (Méthodologique)*, 51, 261-269.
- Meeden, G. (1995). Estimation de la médiane à l'aide d'informations supplémentaires. *Techniques d'enquête*, 21, 81-88.
- Ren, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. *Actes des Journées de méthodologie statistique, INSEE Méthodes*, Tome 1, 100, 263-289.
- Ren, R., et Deville, J.C. (2000). Une généralisation du calage: calage sur les rangs et le calage sur les moments. *II^{ème} Colloque Francophonie sur les Sondages*. Bruxelles.
- Rueda, M.M., Arcos A. et Martínez, M.D. (2003). Difference estimators of quantiles in finite populations. *Test*, 12, 481-496.
- Samdal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.
- Singh, A.C., et Moh, C.A. (1996). Comprendre les estimateurs de calage dans les enquêtes par échantillonnage. *Techniques d'enquête*, 22, 107-116.
- Thompson, M. (1997). *Theory of Sample Surveys*. Chapman & Hall, New York.
- Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 625-646.
- Wu, C., et Sitter, R.R. (2001). Variance estimation for the finite population distribution function with complete auxiliary information. *The Canadian Journal of Statistics*, 29, 289-308.

5. Conclusion

Nous avons élaboré des estimateurs des quantiles fondés sur le paradigme du calage. Ces estimateurs sont particulièrement faciles à appliquer et à interpréter, puisqu'ils sont basés sur les pondérations et les contraintes de calage. De surcroît, il nécessite uniquement la connaissance des quantiles de population des variables auxiliaires, qui peuvent être vectorielles. Lorsqu'on adopte la métrique quadratique, il est possible d'obtenir, pour les poids calés, ainsi que pour les estimateurs de la variance, des expressions analytiques semblables à celles établies pour les estimateurs par calage de totaux. Sur le plan pratique, un aspect intéressant de la nouvelle méthodologie est que les estimateurs proposés sont faciles à calculer; il suffit de transformer les variables auxiliaires, puis d'utiliser les logiciels existants pour calculer les estimateurs par calage.

Au moyen d'une petite étude par simulation, nous avons comparé l'estimateur par calage des quantiles sous la même quantile d'information. L'estimateur fondé sur un modèle, dans lequel est intégré beaucoup plus d'information au sujet des variables auxiliaires, semble préférable sous échantillonnage aléatoire simple et un modèle spécifique correctement, mais est surpassé par le nouvel estimateur lorsque les probabilités d'inclusion de premier ordre sont inégales. En général, l'estimateur proposé se compare favorablement aux estimateurs fondés sur le plan de sondage de Rao et coll. (1990).

Bien que nous soyons concentrés ici sur l'estimation des quantiles par calage sur des quantiles de population connus pour les variables auxiliaires, les estimateurs par calage peuvent être étendus à d'autres problèmes d'estimation importants présentant un intérêt dans le domaine des sondages. Les formulations de ces problèmes mènent toutes à des variables transformées différentes, que nous avons notées z_k dans le présent article. Par exemple, il est possible de formuler un problème de calage pour le coefficient de Gini bien connu, puis de montrer que la solution de ce problème de calage donnera des poids analogues à ceux dérivés ici; cependant, ces poids ne peuvent être déterminés que numériquement. Les travaux devront se poursuivre dans cette direction, afin d'étendre les estimateurs par calage à un cadre plus général, qui inclurait les totaux, les quantiles et les coefficients de Gini en tant que cas particuliers. Un autre domaine de recherche intéressant est celui du choix de l'estimateur de la fonction de répartition. Dans le présent article, nous avons préconisé un estimateur de cette fonction

Remerciements

Les auteurs remercient deux examinateurs anonymes de leurs suggestions et commentaires constructifs, qui leur ont permis d'améliorer considérablement l'article. Ils remercient également Raymond Chambers, Christian Léger, Eric Rancourt, Ulrich Rendel et les participants à la XXXII^e assemblée de la Société statistique du Canada et à la Joint Statistical Meeting de 2004 de leurs discussions et commentaires. Le premier auteur a bénéficié d'une bourse de l'Office allemand d'échanges universitaires (DAAD) et le deuxième, de bourses du Conseil de recherches en sciences naturelles et en génie du Canada et du Fonds québécois de la recherche sur la nature et les technologies du Québec (Canada).

Bibliographie

- Altman, N., et Léger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46, 195-214.
- Bellhouse, D.R., et Stafford, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.
- Cassell, C.M., Samdal, C.-E. et Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika*, 63, 615-620.
- Chambers, R.L., Dorfman, A.H. et Hall, P. (1992). Properties of estimators of finite population distribution functions. *Biometrika*, 79, 577-582.
- Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.
- Chen, J., et Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective use of auxiliary information. *Biometrika*, 80, 107-116.
- Deville, J.-C. (1988). Estimation linéaire et redressement sur l'information auxiliaire d'enquêtes par sondage. Dans *Essais en l'Honneur d'Edmont Malinvaud*. (Éds, A. Monfort, et J.J. Laffond). *Economica*, Paris, 915-929.

Résultats des simulations de Monte Carlo pour l'échantillonnage de la population SLIP982, sous un plan d'échantillonnage PO et la première règle pour la construction des π_k , $k \in U$. Nombre de répliques fixé à $K = 500$

Tableau 7

α	Estimateur	BR _{MC}	V_{MC}	EQM _{MC}	TC	BR _{MC}	V_{MC}	EQM _{MC}	TC
0,25	$\hat{Q}_{y,cal}, \alpha$	0,1393	4,8403	0,956	0,1603	2,8293	2,8493	0,922	0,922
	$\hat{Q}_{y,HT}, \alpha$	-0,0477	5,8276	5,8182	0,934	-0,0227	3,5939	3,5872	0,924
	$\hat{Q}_{y,ra}, \alpha$	0,1648	9,5171	9,5252	0,980	0,1263	4,8687	4,8749	0,972
	$\hat{Q}_{y,diff}, \alpha$	-0,1418	4,7045	4,7152	0,960	-0,0464	2,9213	2,9176	0,936
	$\hat{Q}_{y,CD}, \alpha$	3,9150	3,5279	18,8477	0,584	3,9114	1,9163	17,2112	0,194
0,5	$\hat{Q}_{y,cal}, \alpha$	-0,1746	8,2437	8,2577	0,944	-0,2413	3,6477	3,6986	0,940
	$\hat{Q}_{y,HT}, \alpha$	-0,2824	10,1117	10,1712	0,916	-0,3343	4,5023	4,6050	0,936
	$\hat{Q}_{y,ra}, \alpha$	0,6558	50,4938	50,8228	0,944	0,4263	26,5883	26,7169	0,948
	$\hat{Q}_{y,diff}, \alpha$	-0,5975	17,0315	17,3544	0,972	-0,3496	8,9060	9,0104	0,970
	$\hat{Q}_{y,CD}, \alpha$	4,3173	4,4061	23,0363	0,484	4,0937	2,0711	18,8252	0,184
0,75	$\hat{Q}_{y,cal}, \alpha$	-0,2229	12,1861	12,2114	0,942	-0,2113	6,5823	6,6138	0,952
	$\hat{Q}_{y,HT}, \alpha$	-0,4150	14,2935	14,4371	0,934	-0,2786	7,6597	7,7220	0,934
	$\hat{Q}_{y,ra}, \alpha$	0,7861	47,3844	47,9077	0,980	-0,1344	19,5992	19,5781	0,958
	$\hat{Q}_{y,diff}, \alpha$	0,4347	52,3845	52,4687	0,972	-0,3409	23,8277	23,8962	0,958
	$\hat{Q}_{y,CD}, \alpha$	4,4114	7,7023	27,1478	0,654	4,3549	4,1566	23,1136	0,392

$n = 200$

$n = 100$

Résultats des simulations de Monte Carlo pour l'échantillonnage de la population SLIP982, sous un plan d'échantillonnage PO et la deuxième règle pour la construction des π_k , $k \in U$. Nombre de répliques fixé à $K = 500$

Tableau 8

α	Estimateur	BR _{MC}	V_{MC}	EQM _{MC}	TC	BR _{MC}	V_{MC}	EQM _{MC}	TC
0,25	$\hat{Q}_{y,cal}, \alpha$	0,2392	3,4402	3,4906	0,962	0,1674	1,5214	1,5464	0,952
	$\hat{Q}_{y,HT}, \alpha$	0,0267	4,0027	3,9954	0,940	-0,0370	1,6995	1,6975	0,958
	$\hat{Q}_{y,ra}, \alpha$	0,4402	7,4350	7,6139	0,970	0,1850	3,0687	3,0968	0,978
	$\hat{Q}_{y,diff}, \alpha$	0,0528	3,2842	3,2804	0,972	-0,0127	1,4718	1,4690	0,964
	$\hat{Q}_{y,CD}, \alpha$	2,1458	3,0460	7,6444	0,876	1,9785	1,3010	5,2130	0,690
0,5	$\hat{Q}_{y,cal}, \alpha$	-0,1410	6,5627	6,5695	0,942	-0,2850	2,9662	3,0415	0,954
	$\hat{Q}_{y,HT}, \alpha$	-0,2133	7,6604	7,6906	0,928	-0,2876	3,6017	3,6772	0,926
	$\hat{Q}_{y,ra}, \alpha$	1,0245	43,2773	44,2402	0,930	-0,3075	17,7242	17,7833	0,948
	$\hat{Q}_{y,diff}, \alpha$	-0,1973	14,5261	14,5360	0,958	-0,6111	6,2988	6,6596	0,978
	$\hat{Q}_{y,CD}, \alpha$	2,2140	4,5617	9,4543	0,834	1,8882	2,0393	5,6005	0,738
0,75	$\hat{Q}_{y,cal}, \alpha$	-0,1985	12,6334	12,6476	0,952	-0,0022	5,6442	5,6329	0,966
	$\hat{Q}_{y,HT}, \alpha$	-0,4012	13,5045	13,6384	0,922	-0,1078	6,2239	6,2231	0,934
	$\hat{Q}_{y,ra}, \alpha$	0,7968	44,0650	44,6118	0,958	0,3727	19,1830	19,2836	0,960
	$\hat{Q}_{y,diff}, \alpha$	0,4613	49,6620	49,7755	0,960	0,2340	22,1292	22,1397	0,966
	$\hat{Q}_{y,CD}, \alpha$	2,6329	9,6723	16,5850	0,854	2,6729	4,1179	11,2541	0,738

$n = 200$

$n = 100$

Tableau 5
Résultats des simulations de Monte Carlo pour l'échantillonnage de la population MU284, $y = \text{RMT85}$, $x = \text{REV84}$, sous un plan d'échantillonnage EAS. Nombre de répétitions fixé à $K = 500$

α	Estimateur	B_{MC}	V_{MC}	EQM_{MC}	TC	B_{MC}	V_{MC}	EQM_{MC}	TC
0,25	$\hat{Q}_{y,cal}, \alpha$	1,0161	51,5421	52,4714	0,892	0,6499	24,0662	24,4404	0,954
	$\hat{Q}_{y,HT}, \alpha$	0,3733	110,2572	110,1760	0,960	0,3383	47,2921	47,3120	0,962
	$\hat{Q}_{y,ra}, \alpha$	3,0025	65,4135	74,2979	0,998	2,3856	30,7284	36,3580	0,992
	$\hat{Q}_{y,dif}, \alpha$	2,952	107,7891	114,3084	0,994	2,4083	55,6977	61,3862	0,986
	$\hat{Q}_{y,CD}, \alpha$	-16,5165	1661,0257	1930,4983	0,990	-17,3217	820,7447	1119,1443	0,960
0,5	$\hat{Q}_{y,cal}, \alpha$	-1,6219	215,0336	217,2330	0,870	-0,3419	118,2125	118,0930	0,922
	$\hat{Q}_{y,HT}, \alpha$	0,0075	763,6236	762,0964	0,910	-0,3977	331,2357	330,7314	0,914
	$\hat{Q}_{y,ra}, \alpha$	0,7712	212,8298	212,9988	0,996	-0,2810	136,4382	136,2443	0,996
	$\hat{Q}_{y,dif}, \alpha$	0,3415	283,6718	283,2210	0,998	-1,0104	201,3707	201,9889	0,998
	$\hat{Q}_{y,CD}, \alpha$	17,6124	190,0045	499,8199	n.a.	13,5037	100,2106	282,3611	0,566
0,75	$\hat{Q}_{y,cal}, \alpha$	-5,3477	1023,6924	1050,2431	0,826	-4,7339	443,0660	464,5896	0,926
	$\hat{Q}_{y,HT}, \alpha$	-4,6352	3526,8202	3541,2514	0,938	-5,8890	1242,4858	1274,6812	0,940
	$\hat{Q}_{y,ra}, \alpha$	-1,4390	980,5573	980,6669	0,994	-2,0070	555,5135	558,4305	1,000
	$\hat{Q}_{y,dif}, \alpha$	-5,3988	1464,7867	1491,0041	0,996	-3,9008	744,1604	757,8881	1,000
	$\hat{Q}_{y,CD}, \alpha$	49,3038	2753,8212	5179,1826	n.a.	49,4089	1488,9734	3927,2324	0,596

Tableau 6
Résultats des simulations de Monte Carlo pour l'échantillonnage de la population SLID982, sous un plan d'échantillonnage EAS. Nombre de répétitions fixé à $K = 500$

α	Estimateur	B_{MC}	V_{MC}	EQM_{MC}	TC	B_{MC}	V_{MC}	EQM_{MC}	TC
0,25	$\hat{Q}_{y,cal}, \alpha$	0,1360	3,0390	3,0514	0,956	0,2331	1,6787	1,7297	0,934
	$\hat{Q}_{y,HT}, \alpha$	-0,0596	3,6099	3,6062	0,946	0,0499	1,9277	1,9263	0,918
	$\hat{Q}_{y,ra}, \alpha$	0,3067	6,8815	6,9618	0,970	0,0910	3,0743	3,0764	0,958
	$\hat{Q}_{y,dif}, \alpha$	-0,0504	2,9691	2,9657	0,980	0,0198	1,6139	1,6111	0,952
	$\hat{Q}_{y,CD}, \alpha$	1,1042	2,1180	3,3329	0,922	1,1392	1,2937	2,5888	0,826
0,5	$\hat{Q}_{y,cal}, \alpha$	-0,4034	6,3364	6,4865	0,966	-0,1402	2,9940	3,0076	0,940
	$\hat{Q}_{y,HT}, \alpha$	-0,4157	7,4589	7,6168	0,918	-0,1894	3,5865	3,6151	0,928
	$\hat{Q}_{y,ra}, \alpha$	0,7015	41,8314	42,2399	0,958	0,2238	18,7005	18,7131	0,952
	$\hat{Q}_{y,dif}, \alpha$	-0,4859	14,2083	14,4160	0,970	-0,2740	6,6184	6,6803	0,974
	$\hat{Q}_{y,CD}, \alpha$	0,5702	3,5420	3,8601	0,952	0,6697	1,7559	2,2009	0,932
0,75	$\hat{Q}_{y,cal}, \alpha$	-0,4164	12,4657	12,6142	0,952	-0,2384	5,9118	5,9568	0,950
	$\hat{Q}_{y,HT}, \alpha$	-0,5913	12,5456	12,8701	0,930	-0,3519	6,5496	6,6603	0,926
	$\hat{Q}_{y,ra}, \alpha$	0,7404	48,6836	49,1345	0,954	0,2967	18,5786	18,6294	0,966
	$\hat{Q}_{y,dif}, \alpha$	0,3288	53,6456	53,6464	0,954	0,1841	21,7552	21,7456	0,966
	$\hat{Q}_{y,CD}, \alpha$	0,5966	8,3416	8,6809	0,954	0,5413	4,3692	4,6535	0,936

Tableau 3
Résultats des simulations de Monte Carlo pour l'échantillonnage EAS. Nombre de répétitions fixé à $K = 500$
sous un plan d'échantillonnage de la population MU284, $y = P85$, $x = P75$.

α	Estimateur	B_{MC}	V_{MC}	EQM_{MC}	TC	B_{MC}	V_{MC}	EQM_{MC}	TC
$n = 25$									
0,25	$\hat{Q}_{y, cal., \alpha}$	-0,0343	0,5075	0,5077	0,886	-0,0499	0,2437	0,2457	0,828
	$\hat{Q}_{y, HT., \alpha}$	-0,0266	2,3196	2,3157	0,952	0,0035	1,1087	1,1065	0,936
	$\hat{Q}_{y, ra., \alpha}$	-0,1444	0,3869	0,4070	1,000	-0,0774	0,1684	0,1741	1,000
	$\hat{Q}_{y, diff., \alpha}$	-0,1486	0,3901	0,4114	1,000	-0,0734	0,1723	0,1774	1,000
	$\hat{Q}_{y, CD., \alpha}$	0,4855	0,2791	0,5143	0,906	0,5485	0,1981	0,4985	0,824
0,5	$\hat{Q}_{y, cal., \alpha}$	-0,2762	1,6499	1,7229	0,918	-0,2835	0,9585	1,0370	0,944
	$\hat{Q}_{y, HT., \alpha}$	0,2605	12,5161	12,5589	0,922	-0,0064	5,8466	5,8349	0,916
	$\hat{Q}_{y, ra., \alpha}$	-0,2586	0,8828	0,9479	1,000	-0,4296	0,6701	0,8533	1,000
	$\hat{Q}_{y, diff., \alpha}$	-0,2775	0,9898	1,0648	1,000	-0,4331	0,7492	0,9352	1,000
	$\hat{Q}_{y, CD., \alpha}$	0,9431	0,4054	1,2940	0,866	0,9884	0,2410	1,2175	0,714
0,75	$\hat{Q}_{y, cal., \alpha}$	-0,6229	3,3241	3,7055	0,614	-0,3661	1,8107	1,9411	0,710
	$\hat{Q}_{y, HT., \alpha}$	-0,1414	53,1951	53,1088	0,948	-0,3692	18,8586	18,9572	0,964
	$\hat{Q}_{y, ra., \alpha}$	-0,7925	3,0021	3,6242	1,000	-1,0004	1,4594	2,4573	1,000
	$\hat{Q}_{y, diff., \alpha}$	-0,8230	3,4379	4,1083	1,000	-1,0396	1,5267	2,6044	1,000
	$\hat{Q}_{y, CD., \alpha}$	0,4343	0,5108	0,6984	0,954	0,4485	0,2618	0,4624	0,974
$n = 50$									
0,25	$\hat{Q}_{y, cal., \alpha}$	-0,0441	0,4886	0,4896	0,888	-0,0169	0,2601	0,2599	0,828
	$\hat{Q}_{y, HT., \alpha}$	-0,1698	2,2825	2,3068	0,936	-0,0384	1,1828	1,1819	0,928
	$\hat{Q}_{y, ra., \alpha}$	-0,1509	0,3857	0,4076	1,000	-0,0913	0,2100	0,2179	1,000
	$\hat{Q}_{y, diff., \alpha}$	-0,1634	0,3821	0,4080	1,000	-0,0877	0,2149	0,2221	1,000
	$\hat{Q}_{y, CD., \alpha}$	0,6709	0,3310	0,7805	0,896	0,8792	0,1339	0,9066	0,554
0,5	$\hat{Q}_{y, cal., \alpha}$	-0,3610	1,4881	1,6155	0,920	-0,2236	0,8833	0,9863	0,936
	$\hat{Q}_{y, HT., \alpha}$	-0,0612	11,3969	11,3778	0,926	-0,2712	5,2672	5,3302	0,906
	$\hat{Q}_{y, ra., \alpha}$	-0,3735	1,0009	1,1385	1,000	-0,4130	0,5486	0,7181	1,000
	$\hat{Q}_{y, diff., \alpha}$	-0,3962	1,1271	1,2818	1,000	-0,4217	0,5962	0,7729	1,000
	$\hat{Q}_{y, CD., \alpha}$	1,1740	0,4947	1,8719	0,820	1,3297	0,2146	1,9822	0,474
0,75	$\hat{Q}_{y, cal., \alpha}$	-0,6420	2,6605	3,0674	0,608	-0,4476	1,6212	1,8183	0,708
	$\hat{Q}_{y, HT., \alpha}$	-0,6200	51,2934	51,5752	0,956	-0,6632	17,3625	17,7677	0,966
	$\hat{Q}_{y, ra., \alpha}$	-0,8686	2,8841	3,6329	1,000	-0,9683	1,6494	2,5837	1,000
	$\hat{Q}_{y, diff., \alpha}$	-0,9025	2,9826	3,7911	1,000	-1,0177	1,6340	2,6665	1,000
	$\hat{Q}_{y, CD., \alpha}$	0,4620	0,4501	0,6627	0,982	0,5388	0,2329	0,5228	0,980

Tableau 4
Résultats des simulations de Monte Carlo pour l'échantillonnage de la population MU284, $y = P85$, $x = P75$,
sous un plan d'échantillonnage PO. Nombre de répétitions fixé à $K = 500$

α	Estimateur	B_{MC}	V_{MC}	EQM_{MC}	TC	B_{MC}	V_{MC}	EQM_{MC}	TC
$n = 25$									
0,25	$\hat{Q}_{y, cal., \alpha}$	-0,0441	0,4886	0,4896	0,888	-0,0169	0,2601	0,2599	0,828
	$\hat{Q}_{y, HT., \alpha}$	-0,1698	2,2825	2,3068	0,936	-0,0384	1,1828	1,1819	0,928
	$\hat{Q}_{y, ra., \alpha}$	-0,1509	0,3857	0,4076	1,000	-0,0913	0,2100	0,2179	1,000
	$\hat{Q}_{y, diff., \alpha}$	-0,1634	0,3821	0,4080	1,000	-0,0877	0,2149	0,2221	1,000
	$\hat{Q}_{y, CD., \alpha}$	0,6709	0,3310	0,7805	0,896	0,8792	0,1339	0,9066	0,554
0,5	$\hat{Q}_{y, cal., \alpha}$	-0,3610	1,4881	1,6155	0,920	-0,2236	0,8833	0,9863	0,936
	$\hat{Q}_{y, HT., \alpha}$	-0,0612	11,3969	11,3778	0,926	-0,2712	5,2672	5,3302	0,906
	$\hat{Q}_{y, ra., \alpha}$	-0,3735	1,0009	1,1385	1,000	-0,4130	0,5486	0,7181	1,000
	$\hat{Q}_{y, diff., \alpha}$	-0,3962	1,1271	1,2818	1,000	-0,4217	0,5962	0,7729	1,000
	$\hat{Q}_{y, CD., \alpha}$	1,1740	0,4947	1,8719	0,820	1,3297	0,2146	1,9822	0,474
0,75	$\hat{Q}_{y, cal., \alpha}$	-0,6420	2,6605	3,0674	0,608	-0,4476	1,6212	1,8183	0,708
	$\hat{Q}_{y, HT., \alpha}$	-0,6200	51,2934	51,5752	0,956	-0,6632	17,3625	17,7677	0,966
	$\hat{Q}_{y, ra., \alpha}$	-0,8686	2,8841	3,6329	1,000	-0,9683	1,6494	2,5837	1,000
	$\hat{Q}_{y, diff., \alpha}$	-0,9025	2,9826	3,7911	1,000	-1,0177	1,6340	2,6665	1,000
	$\hat{Q}_{y, CD., \alpha}$	0,4620	0,4501	0,6627	0,982	0,5388	0,2329	0,5228	0,980

4.4 Discussion des résultats empiriques

Nous pensons qu'un modèle tenant compte de l'hétéroscédasticité pourrait améliorer les propriétés de l'estimateur fondé sur un modèle. Cela met en relief le fait que, pour que l'efficacité des estimateurs fondés sur un modèle soit grande, le modèle doit être spécifié correctement.

Les résultats des tableaux 6 à 8 ont trait à la population SLJD982, sous des plans d'échantillonnage EAS et PO avec deux règles pour la détermination des probabilités π_k . Tous les estimateurs présentés au tableau 6 donnent des estimations raisonnablement bonnes du premier quartile et de la médiane, sauf l'estimateur par le ratio $\hat{Q}_{y,ra,\alpha}$, qui est le moins efficace. Le fait que la relation entre les variables dépendante et indépendante ne corresponde pas exactement à un modèle linéaire pourrait expliquer partiellement la performance médiocre de l'estimateur par le ratio dans ce cas. La relation entre x et y n'est pas proportionnelle, si bien que l'estimateur par la différence $\hat{Q}_{y,diff,\alpha}$ semble être préférable à $\hat{Q}_{y,ra,\alpha}$. Cependant, pour $\alpha = 0,75$, ces estimateurs affichent l'EQM la plus élevée, étant tous deux les moins efficaces. Curieusement, dans cette partie de l'expérience, $\hat{Q}_{y,cal,\alpha}$ est supérieur aux estimateurs fondés sur le plan de sondage en ce qui concerne l'EQM. Toutefois, si α est petit, $\hat{Q}_{y,diff,\alpha}$ et $\hat{Q}_{y,cal,\alpha}$ donnent des résultats similaires. Notons que, pour un échantillon de plus grande taille, $\hat{Q}_{y,cal,\alpha}$ et $\hat{Q}_{y,CD,\alpha}$ sont les plus efficaces pour la médiane et le troisième quartile. En fait, l'estimateur fondé sur un modèle $\hat{Q}_{y,CD,\alpha}$ surpasse légèrement $\hat{Q}_{y,cal,\alpha}$, mais il faut souligner qu'il utilise plus d'information auxiliaire.

Les tableaux 7 et 8 présentent les résultats sous les plans d'échantillonnage de Poisson. En général, les estimateurs fondés sur le plan de sondage donnent des résultats comparables à ceux obtenus sous un plan d'échantillonnage aléatoire simple (EAS). Par contre, il n'en est pas ainsi de l'estimateur fondé sur un modèle, qui est moins efficace, vraisemblablement parce qu'il n'intègre pas d'information au sujet du plan d'échantillonnage. Plus précisément, le tableau 7 présente les résultats des simulations sous échantillonnage PO, en utilisant la première règle pour les π_k , $k \in U$. Les taux de couverture de l'estimateur fondé sur un modèle sont particulièrement décevants dans cette expérience, les composantes de biais sont trop importantes dans l'EQM. Les estimateurs fondés sur le plan de sondage fournissent des taux de couverture empiriques nettement plus proches du niveau de confiance nominal de 95 %. Pour les valeurs moyennes et grandes de α , $\hat{Q}_{y,cal,\alpha}$ est l'estimateur le plus efficace. En fait, l'estimateur par calage $\hat{Q}_{y,cal,\alpha}$ donne de bons résultats dans cette expérience. Enfin, le tableau 8 contient les résultats obtenus sous échantillonnage PO avec la deuxième règle pour les π_k . Dans ce cas, $\hat{Q}_{y,ra,\alpha}$ est l'estimateur le moins efficace pour le premier quartile et la médiane, et $\hat{Q}_{y,diff,\alpha}$ est le moins efficace pour $\alpha = 0,75$. En général, $\hat{Q}_{y,cal,\alpha}$ est supérieur aux autres estimateurs dans cette situation, offrant la plus grande efficacité.

Les résultats sont présentés aux tableaux 3 à 8. Nous commençons par discuter des résultats exposés aux tableaux 3 et 4, obtenus en échantillonnant la population MJU84 selon un plan aléatoire simple EAS ou un plan de Poisson PO. Comme on peut le voir, tous les estimateurs ont le même comportement dans les deux études. L'estimateur fondé sur un modèle $\hat{Q}_{y,CD,\alpha}$ semble être le plus efficace parmi ceux analysés lors de l'examen du cas $\alpha = 0,75$ et est, en général, très efficace. Nous nous attendions à ce résultat, puisque la relation entre $x = P85$ et $y = P85$ est fortement linéaire et que l'estimateur fondé sur le modèle est basé sur un modèle hypothétique de régression simple. Par contre, pour $\alpha = 0,25$, les différences d'efficacité sont prononcées par rapport aux autres estimateurs fondés sur des données auxiliaires. Ceux n'utilisant que $\hat{Q}_{x,\alpha}$ comme information sur la variable auxiliaire produisent des résultats assez semblables. Lorsque la taille d'échantillon est petite, les taux de couverture s'écartent habituellement du niveau nominal de 95 %, surtout ceux de $\hat{Q}_{y,cal,\alpha}$, qui sont quelque peu sous-estimés. Cependant, nous observons une certaine amélioration pour $n = 50$, ce qui témoigne de la cohérence des procédures étudiées. Par ailleurs, les taux de couverture de $\hat{Q}_{y,diff,\alpha}$ et $\hat{Q}_{y,ra,\alpha}$ sont toujours égaux à un, ce qui donne à penser que les variances de ces estimateurs sont surestimées. À cause d'une composante de biais importante dans l'EQM, les taux de couverture de l'estimateur fondé sur un modèle se détériorent parfois à mesure qu'augmente la taille de l'échantillon. Nous obtenons les meilleurs taux de couverture en utilisant l'estimateur HT simple, $\hat{Q}_{y,HT,\alpha}$, qui est cependant moins efficace que les autres.

Le tableau 5 donne les résultats pour la deuxième population MJU84, avec $y = RMT85$ et $x = REV84$. La figure 3 semble révéler un phénomène d'hétéroscédasticité dans cette population. Par conséquent, l'utilisation de l'estimateur par le ratio est justifiée puisque la population sous-jacente manifeste ce genre de comportement, il n'est pas étonnant que l'estimateur par le ratio $\hat{Q}_{y,ra,\alpha}$ donne de bons résultats dans cette situation particulière; il surpasse $\hat{Q}_{y,diff,\alpha}$ dans plusieurs cas. Pour toutes les tailles d'échantillon, l'estimateur par le ratio se comporte généralement mieux que $\hat{Q}_{y,cal,\alpha}$. Cependant, pour $n = 50$, l'estimateur par calage semble donner des résultats aussi bons ou légèrement meilleurs que l'estimateur par le ratio. Dans cette expérience, le biais et la variance de l'estimateur fondé sur un modèle $\hat{Q}_{y,CD,\alpha}$ font augmenter sensiblement l'EQM. En outre, dans certains cas, nous n'avons pas pu obtenir les intervalles de confiance pour cet estimateur, car la méthode de Woodruff ne convient pas lorsque la variance est extrêmement grande (l'intervalle de Woodruff devient trop grand et l'hypothèse de linéarité de la fonction de répartition dans cet intervalle n'est plus vérifiée).

scalaire, c'est-à-dire $J = 1$. Des estimateurs de la variance valides de (22) et (23) sont donnés par :

$$V(\hat{\bar{Q}}_{y,ra,\alpha}) = V(\hat{\bar{Q}}_{y,HT,\alpha}) + \left(\frac{\bar{Q}_{y,HT,\alpha}}{\bar{Q}_{x,HT,\alpha}} \right)^2 V(\hat{\bar{Q}}_{x,HT,\alpha}) - 2 \frac{\bar{Q}_{y,HT,\alpha}}{\bar{Q}_{x,HT,\alpha}} C(\hat{\bar{Q}}_{y,HT,\alpha}, \hat{\bar{Q}}_{x,HT,\alpha}),$$

$$V(\hat{\bar{Q}}_{y,diff,\alpha}) = V(\hat{\bar{Q}}_{y,HT,\alpha}) + R^2 V(\hat{\bar{Q}}_{x,HT,\alpha}) - 2R C(\hat{\bar{Q}}_{y,HT,\alpha}, \hat{\bar{Q}}_{x,HT,\alpha}).$$

Ces estimateurs s'appuient sur la variance de $\hat{\bar{Q}}_{y,HT,\alpha}$, ainsi que sur la covariance de $\hat{\bar{Q}}_{y,HT,\alpha}$ et de $\hat{\bar{Q}}_{x,HT,\alpha}$, qui sont estimées selon l'approche de Woodruff (1952) :

$$V(\hat{\bar{Q}}_{y,HT,\alpha}) = \frac{W_y^2}{4z_{1-y/2}^2},$$

$$C(\hat{\bar{Q}}_{y,HT,\alpha}, \hat{\bar{Q}}_{x,HT,\alpha}) = \frac{W_y W_x C(F_y^x(\hat{\bar{Q}}_{x,\alpha}), F_y^x(\hat{\bar{Q}}_{y,\alpha}))}{4z_{1-y/2}^2 [V(F_y^x(\hat{\bar{Q}}_{x,\alpha})) \{V(F_y^x(\hat{\bar{Q}}_{y,\alpha}))\}]^{1/2}},$$

où $W_y = F_y^{-1}(\hat{c}_y) - F_y^{-1}(\hat{c}_y^x)$ et $W_x = F_x^{-1}(\hat{c}_x^x) - F_x^{-1}(\hat{c}_x)$ représentent les intervalles de Woodruff associés à y et à x , avec \hat{c}_y^x et \hat{c}_x^x définis par (18) et (19), $\hat{c}_y^x = \alpha - z_{1-y/2} [V(F_y^x(\hat{\bar{Q}}_{x,\alpha}))]^{1/2}$, $\hat{c}_x^x = \alpha + z_{1-y/2} [V(F_x^x(\hat{\bar{Q}}_{x,\alpha}))]^{1/2}$ et

$$N^{-2} \sum_{i=1}^N \sum_{j=1}^N \Delta_{ij} \left\{ \frac{\pi_{Hj}}{H_{y,s}(\hat{\bar{Q}}_{y,HT,\alpha}, y_j^k) - \alpha} \right\} \left\{ \frac{\pi_i}{H_{x,s}(\hat{\bar{Q}}_{x,HT,\alpha}, x_i^l) - \alpha} \right\} \pi_i.$$

Brièvement, nous nous attendons à ce que $\hat{\bar{Q}}_{y,CD,\alpha}$ donne de bons résultats quand le modèle linéaire décrit la population adéquatement, ce qui motive la comparaison de la nouvelle méthode à un estimateur fondé sur un modèle. En outre, il semble intéressant d'évaluer $\hat{\bar{Q}}_{y,cal,\alpha}$ et les principales propositions fondées sur le plan de sondage, telles que $\hat{\bar{Q}}_{y,diff,\alpha}$ et $\hat{\bar{Q}}_{y,ra,\alpha}$. Les estimateurs $\hat{\bar{Q}}_{y,cal,\alpha}$, $\hat{\bar{Q}}_{y,diff,\alpha}$ et $\hat{\bar{Q}}_{y,ra,\alpha}$ utilisent $\hat{\bar{Q}}_{x,\alpha}$ uniquement pour améliorer les estimations et tiennent compte du plan d'échantillonnage; il s'agit donc de concurrents naturels. Notons que les divers estimateurs inclus dans notre étude sont élaborés sous diverses hypothèses quant à la dimension du vecteur de variables auxiliaires \mathbf{x} et à la

disponibilité de \mathbf{x}_k . Le tableau 2 donne une comparaison des divers estimateurs décrits à la présente section.

Tableau 2

Comparaison des estimateurs par calage proposés et de certains estimateurs importants des quantiles proposés dans la littérature, en ce qui concerne la dimension J de \mathbf{x} et l'information requise au sujet de \mathbf{x}

Estimateur	Dimension de \mathbf{x}	Information requise sur \mathbf{x}
$\hat{\bar{Q}}_{y,HT,\alpha}$	s.o.	aucune
$\hat{\bar{Q}}_{y,CD,\alpha}$	$J \geq 1$	$\mathbf{x}_k, k \in U/s$
$\hat{\bar{Q}}_{y,ra,\alpha}$	$J = 1$	$\hat{\bar{Q}}_{x,\alpha}$
$\hat{\bar{Q}}_{y,diff,\alpha}$	$J = 1$	$\hat{\bar{Q}}_{x,\alpha}$
$\hat{\bar{Q}}_{y,cal,\alpha}$	$J \geq 1$	$\hat{\bar{Q}}_{x,\alpha}$

4.3 Mesures fréquentistes

Notre objectif est d'évaluer les estimateurs en nous basant sur le biais et la variance. D'autres considérations importantes sont l'erreur quadratique moyenne (EQM) et les taux de couverture des intervalles de confiance.

Soit $\hat{\bar{Q}}_{y,\alpha}$ un estimateur du quantile de population $\bar{Q}_{y,\alpha}$. Supposons que $\hat{\bar{Q}}_{y,\alpha}^{(v)}$ est l'estimateur du quantile calculé en utilisant l'échantillon $v, v = 1, \dots, K$. La moyenne de Monte Carlo E_{MC} , le biais de Monte Carlo B_{MC} et la variance de Monte Carlo V_{MC} sont donnés par les formules usuelles, c'est-à-dire

$$E_{MC}(\hat{\bar{Q}}_{y,\alpha}) = K^{-1} \sum_{v=1}^K \hat{\bar{Q}}_{y,\alpha}^{(v)},$$

$$B_{MC} = E_{MC}(\hat{\bar{Q}}_{y,\alpha}) - \bar{Q}_{y,\alpha},$$

$$V_{MC}(\hat{\bar{Q}}_{y,\alpha}) = K^{-1} \sum_{v=1}^K \{ \hat{\bar{Q}}_{y,\alpha}^{(v)} - E_{MC}(\hat{\bar{Q}}_{y,\alpha}) \}^2.$$

Notre critère principal de détermination de l'efficacité est l'EQM de Monte Carlo, définie par $EQM_{MC} = K^{-1} \sum_{v=1}^K (\hat{\bar{Q}}_{y,\alpha}^{(v)} - \bar{Q}_{y,\alpha})^2$. Nous calculons les intervalles de confiance au niveau de confiance de 95 %, selon un processus décrites aux sections qui précèdent. Pour un estimateur $\hat{\bar{Q}}_{y,\alpha}^{(v)}$ et l'estimateur de sa variance $V^{(v)}$, $v = 1, \dots, K$, les taux de couverture (TC) au niveau de confiance de 95 % sont calculés selon l'expression

$$TC(\hat{\bar{Q}}_{y,\alpha}) = K^{-1} \sum_{v=1}^K I \left\{ \left[\hat{\bar{Q}}_{y,\alpha}^{(v)} - 1.96 \sqrt{V^{(v)}} \right] \leq \left[\hat{\bar{Q}}_{y,\alpha}^{(v)} + 1.96 \sqrt{V^{(v)}} \right] \right\},$$

où $I(\cdot)$ est la fonction indicateur de l'ensemble \mathcal{A} . Les taux de couverture sont donnés dans la colonne intitulée TC. Rappelons que nous adoptons $K = 500$ pour toutes les études.

$\alpha = 0,75$, en plus de la médiane et du premier quartile. À la section suivante, nous décrirons les estimateurs utilisés dans l'étude.

4.2 Estimateurs inclus dans l'étude empirique

Puisque l'un de nos objectifs est de proposer des estimateurs ayant des propriétés raisonnables en ce qui concerne le biais, la variance et les taux de couverture des intervalles de confiance, nous comparons le nouvel estimateur défini par (9) fondé sur la métrique (2) à certains estimateurs des quantiles populaires proposés dans la littérature.

Pour commencer, nous incluons l'estimateur fondé sur le plan de sondage simple basé sur l'inversion de l'estimateur $F_Y(t) = \sum_{k \in U} d_k H_{Y,s}(t, Y_k) / \sum_{k \in U} d_k$:
 L'estimateur (17) n'utilise pas d'information auxiliaire. Un estimateur possible de la variance est

$$\hat{Q}_{Y,HT,\alpha} = F_Y^{-1}(\alpha). \quad (17)$$

estimateur possible de la variance est

$$V\{F_Y(\hat{Q}_{Y,\alpha})\} = \sum_{s=1}^S \sum_{k \in U/s} \Delta_{sk} \left\{ \frac{\pi_k}{H_{Y,s}(\hat{Q}_{Y,HT,\alpha}, Y_k) - \alpha} \right\} \left\{ \frac{\pi_l}{H_{Y,s}(\hat{Q}_{Y,HT,\alpha}, Y_l) - \alpha} \right\} \frac{\pi_l}{\pi_k},$$

où $N = \sum_{k \in U} d_k$, et les intervalles de confiance peuvent être calculés au moyen de

$$[F_Y^{-1}(\hat{Q}_{Y,\alpha}), F_Y^{-1}(\hat{Q}_{Y,\alpha})],$$

où

$$\hat{Q}_{Y,\alpha} = \alpha - z_{1-\alpha/2} [V\{F_Y(\hat{Q}_{Y,\alpha})\}]^{1/2}, \quad (18)$$

$$\hat{Q}_{Y,\alpha} = \alpha + z_{1-\alpha/2} [V\{F_Y(\hat{Q}_{Y,\alpha})\}]^{1/2}. \quad (19)$$

Pour plus de détails, consulter Samdal et coll. (1992, page 202).

Nous incluons également dans notre étude empirique l'estimateur fondé sur un modèle de Chambers et Dunsan (1986), qui est motivé par un modèle de superpopulation linéaire $Y_k = \beta_0 + \beta' X_k + e_k$, $k \in U$, où e_k forme une suite de variables aléatoires indépendantes et de même loi de moyenne nulle et de variance finie. Leur estimateur est défini par

$$\hat{Q}_{Y,CD,\alpha} = \inf\{t | F_{Y,CD}(t) \geq \alpha\}, \quad (20)$$

où $F_{Y,CD}(t) = N^{-1} \{ \sum_{k \in U} H(t - Y_k) + \sum_{k \in U/s} \hat{Q}(t - \hat{Y}_k) \}$ représente un estimateur fondé sur un modèle de la fonction de répartition,

$$G(n) = n^{-1} \sum_{k \in U} H(n - \hat{e}_k) \quad (21)$$

représente la fonction de répartition empirique des résidus pond aux prédictions selon les moindres carrés. Puisque $\hat{e}_k = Y_k - \hat{Y}_k$, $k \in s$, et $\hat{Y}_k = \beta_0 + \beta' X_k$, $k \in U/s$ correspond à l'impute essentiellement la valeur inconnue

La construction d'un intervalle de confiance pour $\hat{Q}_{Y,CD,\alpha}$ repose sur l'estimation de la variance $V\{F_{Y,CD}(t)\}$. Toute-fois, ce problème d'estimation pose des difficultés, puisque toute formule analytique de la variance dépend du modèle hypothétique. En outre, les expressions analytiques de ce genre font intervenir des estimateurs à moyen de la densité, d'une fonction noyau et d'une fenêtre (bandwidth). Pour toutes ces raisons, nous avons décidé d'appliquer les estimateurs de la variance par le jackknife avec suppression d'une unité étudiée par Wu et Sitter (2001), qui ont montré la convergence des estimateurs proposés de la variance. Dans le contexte des sondages, diverses méthodes de rééchantillonnage, y compris le jackknife, sont présentées dans Kovan, Rao et Wu (1988). La technique du jackknife consiste à supprimer une unité et à recalculer l'estimateur. Soit $s_i = s/\{i\}$ l'échantillon sans l'unité i . Considérons β_{0i} et β'_{i1} , les estimateurs par la régression de β_0 et β calculés sur s_i . Sous un modèle de régression simple, définissons

$$F_i^* = (n-1)^{-1} \sum_{k \in U/s} \left[N^{-1} \sum_{l \in U/s} H(\hat{Q}_{Y,CD,\alpha} - \beta_{0i}(x_l - x_k) - \beta'_{i1}(Y_l - Y_k)) \right].$$

Un estimateur de la variance convergent de $V\{F_{Y,CD}(\hat{Q}_{Y,CD,\alpha})\}$ est donné par

$$V\{F_{Y,CD}(\hat{Q}_{Y,\alpha})\} = \frac{n}{n-1} \sum_{i=1}^n (F_i^* - F^*)^2$$

$$+ \frac{f(1-f)}{N-n} \sum_{k \in U/s} \{ \hat{Q}(\hat{Q}_{Y,CD,\alpha} - \hat{Y}_k) - \hat{Q}(\hat{Q}_{Y,CD,\alpha} - \hat{Y}_k) \},$$

où $f = n/N$ est la fraction d'échantillonnage, $F^* = N^{-1} \sum_{k \in U} F_k^*$, et \hat{Q} est donné par (21). En partant de $V\{F_{Y,CD}(\hat{Q}_{Y,\alpha})\}$, nous pouvons maintenant calculer les intervalles de confiance pour $\hat{Q}_{Y,\alpha}$ en suivant l'approche de l'inversion.

Puisque notre méthode ne nécessite que la connaissance du vecteur des quantiles $\mathbf{Q}_{X,\alpha}$, nous incluons dans notre étude les estimateurs par le ratio et par la différence des quantiles étudiés dans Rao et coll. (1990) :

$$\hat{Q}_{Y,RA,\alpha} = \hat{Q}_{X,\alpha} (\hat{Q}_{Y,HT,\alpha} / \hat{Q}_{X,HT,\alpha}), \quad (22)$$

$$\hat{Q}_{Y,DIFF,\alpha} = \hat{Q}_{Y,HT,\alpha} + R(\hat{Q}_{X,\alpha} - \hat{Q}_{X,HT,\alpha}), \quad (23)$$

où $\hat{Q}_{Y,HT,\alpha}$ est donné par (17) et $\hat{Q}_{X,HT,\alpha}$ est calculé de la même façon; l'estimateur par le ratio donné par $R = \sum_{k \in U} d_k Y_k / \sum_{k \in U} d_k X_k$ fournit un estimateur convergent de $R = \sum_{k \in U} Y_k / \sum_{k \in U} X_k$. Notons que les estimateurs (22) et (23) sont élaborés en se fondant sur une variable auxiliaire

lequel nous supposons que ces facteurs fournissent des mesures de taille appropriée pour les unités dans les diverses classes âge-sexe (voir, par exemple, Särndal et coll. (1992, page 87)); pour ces unités, plus d'hommes que de femmes sont susceptibles d'être sélectionnés et pour les deux sexes, les adultes de 27 à 37 ans et ceux de 38 à 46 ans sont plus susceptibles que les autres d'être inclus dans l'échantillon.

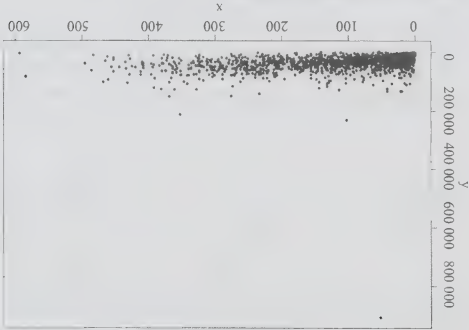


Figure 5. Population SLID982, où la variable dépendante est le revenu impossible et la variable indépendante, la durée de l'emploi courant (en mois).

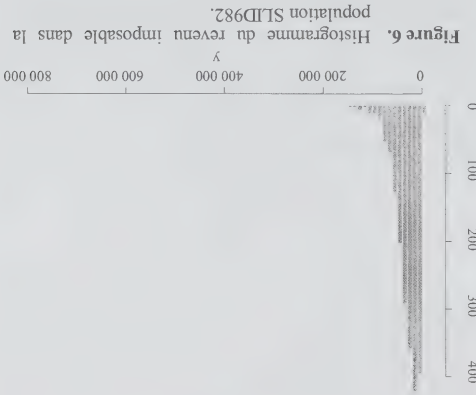


Figure 6. Histogramme du revenu impossible dans la population SLID982.

Tableau 1

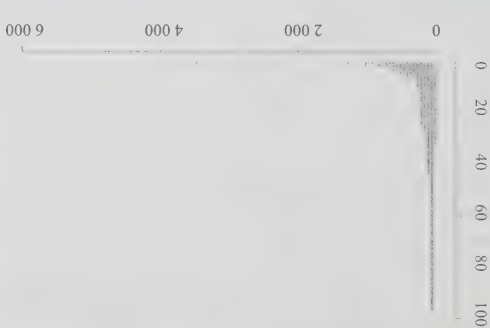
Facteur p_{2k} selon l'âge et le sexe de l'individu k ,

dans la population SLID982

Âge		Sexe	
16 à 25 ans	1	Hommes	3
27 à 37 ans	2	Hommes	6
38 à 46 ans	3	Femmes	1
47 à 69 ans	4	Femmes	2

Dans ces trois études, nous estimons les quartiles, c'est-à-dire les paramètres de population $Q_{\gamma,\alpha}$ pour lesquels $\alpha = 0,25, 0,5$ et $0,75$. Puisque les variables d'intérêt présentent une distribution fortement asymétrique, il semble particulièrement intéressant d'étudier le quartile correspondant à

Figure 4. Histogramme de la variable RMT85 dans la population MU284.



La troisième population est basée sur un sous-échantillon aléatoire de l'Enquête sur la dynamique du travail et du revenu, noté SLID982 (SLID, *Survey of Labour and Income Dynamics*). L'enquête a été réalisée par Statistique Canada en 1998. Pour simplifier, nous n'avons sélectionné que les entrées pour lesquelles aucune valeur ne manquait. La taille d'étude, nous supposons qu'il s'agit d'une population (la taille originale de l'échantillon de cette enquête est d'environ 60 000). Le revenu impossible (en milliers de dollars) est la variable cible, tandis que la durée en mois de l'emploi courant est la variable auxiliaire. Comme l'illustre la figure 5, la relation linéaire entre le revenu impossible et la durée de l'emploi est moins prononcée. Cependant, les deux variables ne semblent pas être indépendantes. À la figure 6, nous voyons que la variable d'intérêt présente un coefficient d'asymétrie élevé. Nous avons tiré 500 échantillons à partir de la population SLID982, selon les plans d'échantillonnage EAS et PO. Nous avons considéré comme taille d'échantillon (taille espérée d'échantillon) $n = 100$ et $n = 200$. Pour l'échantillonnage PO, nous avons défini les probabilités de sélection de premier ordre, $\pi_k, k \in U$, en fonction de deux règles. Sous la première, nous avons créé les π_k de telle sorte qu'elles soient approximativement proportionnelles à la variable d'intérêt, c'est-à-dire le revenu impossible (aux fins de notre étude, nous supposons qu'il est possible de créer de telles π_k). Puisque certaines valeurs de y_k sont négatives dans la population, nous choisissons $p_{1k} = y_k - \min\{y_k, k \in U\} + 1$ et nous définissons $\pi_k = E(n_s)p_{1k} / \sum_{k \in U} p_{1k}$, où $E(n_s)$ représente la taille espérée de l'échantillon, dans notre cas $E(n_s) = 100$ et 200. En vertu de la deuxième règle, les π_k ont été créées proportionnelles aux entrées du tableau 1. Autrement dit, pour chaque $k \in U$, il existe un facteur p_{2k} , qui est déterminé par le groupe âge-sexe de l'individu k . Alors, $\pi_k = E(n_s)p_{2k} / \sum_{k \in U} p_{2k}$, où les facteurs p_{2k} sont donnés au tableau 1. Les facteurs p_{2k} du tableau 1 sont fondés sur un plan d'échantillonnage hypothétique, dans

échantillons finis et de les comparer à celles des estimateurs des quantiles généralement décrits dans la littérature. À la présente section, nous entreprenons des expériences par simulation afin d'illustrer empiriquement les nouveaux estimateurs. Nous nous intéressons surtout à leur biais et à leur variance empirique en population réelle. Nous étudions également les propriétés de couverture des intervalles de confiance, qui représentent aussi une question d'intérêt pratique.

Afin de répondre partiellement à ces questions, nous avons exécuté trois petites études par simulation, dans le cadre desquelles, pour plusieurs plans d'échantillonnage et pour des populations réelles, nous comparons l'estimateur par calage proposé des quantiles aux estimateurs généralement utilisés à l'heure actuelle. À la sous-section 4.1, nous décrivons en détail les populations étudiées et discutons des plans d'échantillonnage choisis. À la sous-section 4.2, nous présentons les estimateurs inclus dans l'étude empirique et, à la sous-section 4.3, nous décrivons les mesures fréquentistes (biais, variance et erreur quadratique moyenne empirique, taux de couverture des intervalles de confiance). Enfin, à la sous-section 4.4, nous analysons nos résultats empiriques.

4.1 Description des populations réelles et des plans d'échantillonnage

Les populations réelles sont représentées aux figures 1 à

6. La première, notée MU284, est tirée de Särndal et coll. (1992, annexe B). Elle est constituée de $N = 284$ municipalités de la Suède. Nous retenons comme variable d'intérêt la population en 1985 (variable P85) et nous supposons que l'information auxiliaire disponible est la population en 1975 (variable P75). Les deux variables sont mesurées en milliers. À la figure 1, la variable P85 est exprimée en fonction de la variable P75; comme prévu, la relation entre P85 et P75 est fortement linéaire. La variable P95 suit une loi hautement asymétrique, comme l'illustre la figure 2. Dans cette population, nous avons tiré 500 échantillons selon un plan d'échantillonnage aléatoire simple sans remise (EAS). En outre, nous avons exécuté la même étude sous un plan d'échantillonnage avec probabilités de sélection inégales, c'est-à-dire le plan d'échantillonnage de Poisson (PO). Les propriétés du plan d'échantillonnage PO sont décrites dans Särndal et coll. (1992). Étant donné la grande fourchette de valeurs de y , nous n'avons pas pu construire des probabilités de sélection d'échantillon π_k de la forme $\pi_k \propto y_k$, car certaines π_k auraient dû être supérieures à l'unité. Aux fins de notre illustration, nous avons déterminé les probabilités de sélection en utilisant la relation $\pi_k \propto 0,2y_k + 0,05$ (nous reconnaissons que les π_k sont idéalisées, puisque y_k n'est pas disponible en pratique). Sous le plan d'échantillonnage EAS (plan d'échantillonnage PO), nous considérons les tailles d'échantillon (tailles

espérées d'échantillon) $n = 25$ et $n = 50$.

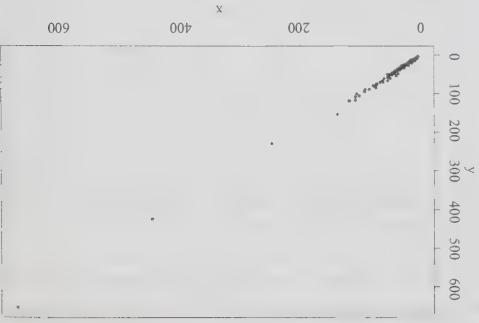


Figure 1. Population MU284, où $y = P85$ et $x = P75$.

Pour la deuxième étude, nous avons choisi la population MU284, mais retenu comme variable d'intérêt $y = RMT85$, qui représente les recettes de l'impôt municipal de 1985 (en millions de couronnes). Ici, la variable auxiliaire choisie est $x = REV84$, qui représente les valeurs immobilières selon les évaluations de 1984 de chaque municipalité (en millions de couronnes). Comme le montre la figure 3, la relation entre x et y est quelque peu étalée pour les grandes valeurs de x . L'histogramme de la variable RMT85 révèle que celle-ci suit une loi asymétrique (figure 4). Pour cette étude, nous avons tiré, selon le plan d'échantillonnage EAS, 500 échantillons de

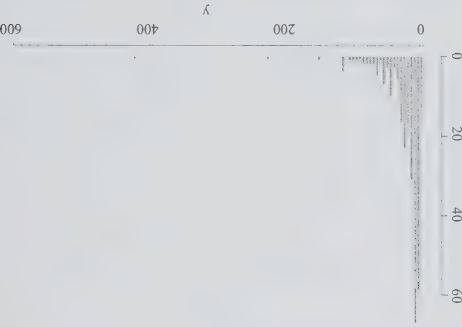


Figure 2. Histogramme de la variable P85 dans la population MU284.

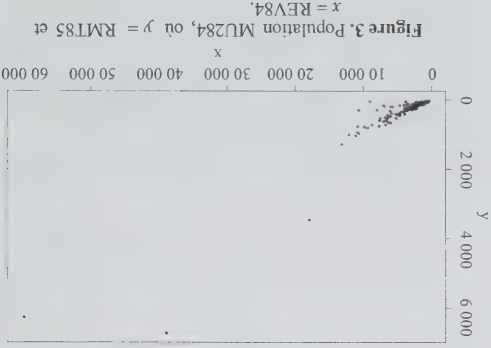


Figure 3. Population MU284, où $y = RMT85$ et $x = REV84$.

Preuve. Afin de prouver la proposition 1, soulignons d'abord que, puisque que la première contrainte $\sum_{j=1}^N w_k = N$ doit être satisfaite, il s'en suit que $F_{x_j, \text{cal}}(t) = N^{-1} \sum_{j=1}^N w_k H_{x_j, s}(t, x_j^k)$. En nous inspirant de Deville et Särndal (1992), nous pouvons montrer que le vecteur $\mathbf{a}_k = (1, a_{1k}, \dots, a_{jk})'$ satisfait

$$\mathbf{a}_k = \left(1, \frac{\partial F_{x_j, \text{cal}}}{\partial w_k}, \dots, \frac{\partial F_{x_j, \text{cal}}}{\partial w_k} \right), \quad (12)$$

que nous évaluons maintenant explicitement. L'évaluation des dérivées nous donne $a_{jk} = N^{-1} H_{x_j, s}(t, x_j^k)$, $j = 1, \dots, J$, évalué à $t = \mathcal{Q}_{x_j, \alpha}$. Ceci mène à

$$a_{jk} = \begin{cases} N^{-1}, & x_j^k \leq L_{x_j, s}(\mathcal{Q}_{x_j, \alpha}), \\ N^{-1} \mathbf{p}_{x_j, s}(\mathcal{Q}_{x_j, \alpha}), & x_j^k = U_{x_j, s}(\mathcal{Q}_{x_j, \alpha}), \\ 0, & x_j^k > U_{x_j, s}(\mathcal{Q}_{x_j, \alpha}). \end{cases}$$

$j = 1, \dots, J$, tel qu'annoncé.

Dans (11), \mathbf{T}_k peut être interprété comme étant la valeur espérée de $\sum_{j=1}^J a_{jk}$. Dans l'estimateur de la fonction de répartition (6), les poids dérivés (10) dépendent des variables \mathbf{a}_k , $k \in s$ définies par (12). Notons qu'elles correspondent à une certaine transformation de la variable auxiliaire \mathbf{x}_k . La différence entre les poids utilisés pour les totaux et pour les quantiles tient à cette variable \mathbf{a}_k . Lorsque \mathbf{a}_k est remplacée par \mathbf{x}_k , nous obtenons les poids originaux pour les totaux. Par conséquent, il est utile d'interpréter cette nouvelle variable. Lors de l'estimation d'un total, l'effet sur la j^{e} contrainte de calage est mesuré par x_{jk} , pour chaque unité $k \in s$. Dans notre cadre, l'effet de l'unité k est maintenant donné par N^{-1} si $x_{jk} \leq L_{x_j, s}(\mathcal{Q}_{x_j, \alpha})$; il correspond au facteur $N^{-1} \mathbf{p}_{x_j, s}(\mathcal{Q}_{x_j, \alpha})$ quand $x_{jk} = U_{x_j, s}(\mathcal{Q}_{x_j, \alpha})$ et est nul ailleurs. À la section 5, nous discuterons d'autres problèmes d'estimation menant à des variables \mathbf{a}_k différentes.

Compte tenu des similarités entre les estimations des totaux et des quantiles, nous pouvons également envisager l'estimation de la variance. Nous abordons cette question à la sous-section suivante.

3.3 Estimation de la variance et intervalles de confiance

Comme nous l'avons décrit à la section précédente, l'estimateur $\hat{\mathcal{Q}}_{x_j, \text{cal}}$ présente plusieurs similarités avec l'estimateur GREG habituel pour les totaux de population. Les variables transformées données par (12) constituent la principale différence entre les estimateurs par calage des quantiles et ceux des totaux. Il se trouve que, étant donné la similarité structurelle avec les estimateurs par calage originaux, il est facile de déterminer un intervalle de confiance

pour l'estimateur proposé $\hat{\mathcal{Q}}_{x_j, \text{cal}, \alpha}$. Nous considérons la construction d'intervalles de confiance suivant l'approche de Woodruff (1952). L'intervalle de confiance est donné au résultat 1.

Résultat 1 (intervalle de confiance de Woodruff pour l'estimateur par calage des quantiles). L'intervalle de confiance basé sur l'approche de Woodruff (1952), lorsqu'on utilise l'estimateur par calage (9) pour le quantile $\mathcal{Q}_{y, \alpha}$, est donné par

$$[\hat{F}_{y, \text{cal}}^{-1}(\hat{c}_{1y}), \hat{F}_{y, \text{cal}}^{-1}(\hat{c}_{2y})], \quad (13)$$

où $\hat{c}_{1y} = \alpha - z_{1-\gamma/2} [F_{y, \text{cal}}(\mathcal{Q}_{y, \alpha})]^{1/2}$ et $\hat{c}_{2y} = \alpha + z_{1-\gamma/2} [F_{y, \text{cal}}(\mathcal{Q}_{y, \alpha})]^{1/2}$. La procédure résultante donne un intervalle de confiance approximatif pour $\mathcal{Q}_{y, \alpha}$ à un niveau de confiance $1 - \gamma$ précisé.

Preuve. En supposant que $F_{y, \text{cal}, \alpha}(\mathcal{Q}_{y, \alpha})$ suit approximativement une loi normale, il s'ensuit que $\Pr(\hat{c}_{1y} \leq F_{y, \text{cal}, \alpha}(\mathcal{Q}_{y, \alpha}) \leq \hat{c}_{2y})$ devrait être approximativement égal à $1 - \gamma$, si l'on choisit

$$c_{1y} = \alpha - z_{1-\gamma/2} [F_{y, \text{cal}}(\mathcal{Q}_{y, \alpha})]^{1/2}, \quad (14)$$

$$c_{2y} = \alpha + z_{1-\gamma/2} [F_{y, \text{cal}}(\mathcal{Q}_{y, \alpha})]^{1/2}, \quad (15)$$

où z_γ représente le γ^{e} quantile de la loi normale standard $N(0, 1)$. Puisque $F_{y, \text{cal}, \alpha}(\mathcal{Q}_{y, \alpha})$ représente essentiellement une moyenne d'échantillon, un estimateur possible de la variance justifié par la linéarisation de Taylor classique est donnée par

$$\Delta_{\hat{y}} \{F_{y, \text{cal}}(\mathcal{Q}_{y, \alpha})\} = N^{-2} \sum_{j=1}^J \sum_{k \in s} w_k e_k (w_j e_j), \quad (16)$$

où $\Delta_{\hat{y}} = \pi_{\hat{y}} - \pi_k \pi_j$; les poids $w_k, k \in s$, correspondent aux poids calés (3) qui se réduisent à (10) lorsque D est la fonction de distance quadratique (2); les résidus sont données par $e_k = \hat{F}_{y, \text{cal}}(\mathcal{Q}_{y, \alpha}) - \mathbf{a}_k' \mathbf{B}$ ou

$$\mathbf{B}_s = \left(\sum_{k \in s} w_k \mathbf{a}_k \mathbf{a}_k' \right)^{-1} \sum_{k \in s} w_k \mathbf{a}_k \mathbf{a}_k' H_{y, s}(\mathcal{Q}_{y, \text{cal}, \alpha}, y^k)$$

représente l'estimateur des coefficients de régression. Puisque les constantes c_{1y} et c_{2y} données par (14) et (15) dépendent de $\{F_{y, \text{cal}}(\mathcal{Q}_{y, \alpha})\}$, nous pouvons les estimer en utilisant l'estimateur de la variance (16).

Dans le résultat 1, soulignons que Deville et Särndal (1992) ont préconisé l'utilisation d'un estimateur de la variance des estimateurs par calage des totaux de population. À la section 4, nous étudions empiriquement les propriétés de l'estimateur par calage proposé (9) et l'intervalle de confiance donné par l'expression (13).

4. Résultats des simulations

D'un point de vue pratique, il est logique d'étudier les propriétés des nouveaux estimateurs par calage pour des

variables auxiliaires; donc, l'efficacité des estimateurs par calage ainsi obtenus devrait, en principe, être plus grande.

Remarque 3

L'estimateur par calage proposé (9) est obtenu par calage sur les quantiles de population. Une autre possibilité a été examinée par Ren (2002) qui a calé les estimateurs sur les moments de population, jusqu'à l'ordre m , d'une même loi. Plus précisément, Ren (2002) a proposé des estimateurs par calage des quantiles satisfaisant des contraintes de la forme $\sum_{j=1}^s w_j x_{j:m}^* = \sum_{j=1}^s U_{j:m}^*$, $m = 0, 1, \dots, M$. Le calage sur divers moments de la même loi est étroitement associé au calage sur divers quantiles de la même variable. Pour d'autres généralisations du paradigme de calage sur les moments, consulter aussi Ren et Deville (2000), ainsi que Harms (2003).

(2003).

3.2. Solution analytique des poids calés quand D est la métrique quadratique

Lorsque l'on adopte la fonction de distance quadratique (2), il est possible de dériver une solution explicite du problème d'optimisation (3). Cette situation est semblable à celle des estimateurs par calage des totaux, où les poids de l'estimateur GREG sont obtenus explicitement sous la métrique (2). Une analyse minutieuse du problème d'estimation dans le cas des quantiles révèle d'importantes similitudes, dues au fait que les estimateurs donnés par (7) sont des sommes pondérées des variables $\{H_{x_j,s}(t, x_{j,k}), k \in s\}$, $j = 1, \dots, J$. Cela est énoncé dans la proposition 1.

Proposition 1 (poids calés pour la métrique quadratique). *Considérons la fonction de distance quadratique (2). Le vecteur de poids w qui résout le problème d'optimisation (3) satisfait la relation :*

$$w_k = d_k^*(1 + q_k a_k^* \lambda_s), k \in s, \quad (10)$$

où le vecteur $\lambda_s = (\lambda_{0,s}, \dots, \lambda_{J,s})'$ est déterminé par la voie des $J+1$ contraintes sous la forme :

$$\lambda_s = \left(\sum_{j=1}^s d_j^* q_j a_j^* \right)^{-1} \left(T_s - \sum_{j=1}^s d_j^* a_j^* \right), \quad (11)$$

avec $T_s = (N, \alpha, \dots, \alpha)'$ et les composantes de $a_k = (1, a_{1k}, \dots, a_{Jk})'$ sont données par

$$a_{jk} = \begin{cases} 0, & x_{jk} > U_{x_j,s}(\hat{O}_{x_j,\alpha}^{cal}), \\ N^{-1} g_{x_j,s}(\hat{O}_{x_j,\alpha}^{cal}), & x_{jk} = U_{x_j,s}(\hat{O}_{x_j,\alpha}^{cal}), \\ N^{-1}, & x_{jk} \leq L_{x_j,s}(\hat{O}_{x_j,\alpha}^{cal}). \end{cases}$$

avec $j = 1, \dots, J$.

que toutes les valeurs x_{jk} figurant dans l'échantillon s , alors $F_{x_j,s}^{cal}(\hat{O}_{x_j,\alpha}^{cal})$ sera égal à zéro ou à un, quels que soient les poids w que l'on choisit. Donc, dans ces cas, il peut arriver que les contraintes de calage ne puissent être satisfaites. Cependant, quand le comportement de l'échantillon diffère considérablement de celui de la population cible, il convient d'examiner tout ajustement d'un tel cas extrême. En pratique, elle se produit rarement, à moins que l'on choisisse une valeur de α très proche de zéro ou de un. Notons qu'il pourrait être impossible d'obtenir une solution si la taille n de l'échantillon est petite. Le cas échéant, nous pourrions considérer le minimum ou le maximum d'échantillon comme un estimateur possible ou recourir au simple estimateur de la fonction de répartition fondé sur le plan de sondage.

Le deuxième problème éventuel est que certains poids w_k pourraient être négatifs. Dans ces conditions, $F_{x_j,s}^{cal}$ n'est plus bijectif, ce qui ne cause pas d'ennui à condition que $F_{x_j,s}^{-1}(\alpha)$ demeure déterminé de façon unique. Nous pouvons éviter ce problème en contraignant tous les poids à des valeurs strictement positives, grâce à l'utilisation d'une métrique appropriée $D(\cdot, \cdot)$. Cette approche a été adoptée par Kováčević (1997) (pour plus de précisions sur les fonctions de distance produisant des poids positifs, voir également Deville et Samdal (1992), ainsi que Singh et Mohl (1996)).

Remarque 1

Les estimateurs proposés des fonctions de répartition (6) et (7) s'appuient sur une interpolation linéaire. Par souci d'uniformisation, la fonction de répartition de la population, qui est aussi une fonction échelon, pourrait également être définie en se fondant sur une interpolation linéaire. En pratique, les deux définitions correspondent à des comportements qui ne diffèrent que légèrement si la population N est suffisamment grande. Cependant, il convient de souligner que, si la taille de la population N est assez faible, l'utilisation d'une interpolation pour définir les fonctions de répartition pourrait valoir la peine.

Remarque 2

Dans le problème d'optimisation (3), nous avons réalisé le calage sur un quantile particulier. Cette approche pourrait être étendue en permettant le calage sur un ensemble fini de quantiles, si ce genre d'information est disponible. Plus précisément, supposons que, pour une variable auxiliaire x , les quantiles α_m dénotés \hat{O}_{x,α_m} , $m = 1, \dots, M$ sont connus, où $M < n - 1$. Dans ce cas, nous pourrions envisager les contraintes de calage $F_{x,\alpha_m}^{cal}(\hat{O}_{x,\alpha_m}^{cal}) = \alpha_m$, $m = 1, \dots, M$ et résoudre le problème d'optimisation (3) avec ces contraintes supplémentaires. Naturellement, cette information produit une description plus complète de la distribution des

$$(7) \quad \hat{F}_{x^{j,s}, \text{cal}}^{\alpha}(t) = \frac{\sum_s w_k H_{x^{j,s}}(t, x^{j,k})}{\sum_s w_k^s},$$

où, dans (4) et (5), la fonction de Heaviside H est remplacée par la fonction légèrement modifiée

$$(8) \quad H_{y^{j,s}}(t, y^k) = \begin{cases} 1, & y^k \leq L_{y^{j,s}}(t), \\ \beta_{y^{j,s}}(t) & y^k = U_{y^{j,s}}(t), \\ 0, & y^k > U_{y^{j,s}}(t), \end{cases}$$

où $L_{y^{j,s}}(t) = \max\{\{y^k, k \in s \mid y^k \leq t\} \cup \{-\infty\}\}$,

$U_{y^{j,s}}(t) = \min\{\{y^k, k \in s \mid y^k > t\} \cup \{\infty\}\}$ et $\beta_{y^{j,s}}(t) = \{t - L_{y^{j,s}}(t)\} / \{U_{y^{j,s}}(t) - L_{y^{j,s}}(t)\}$. La fonction $H_{x^{j,s}}(t, x^k)$ est définie de la même façon. Les estimateurs (6) et (7), basés sur les fonctions $H_{y^{j,s}}(t, y^k)$ et $H_{x^{j,s}}(t, x^k)$, sont appelés estimateurs interpolés pour les fonctions de distribution $F_y(t)$ et $F_x(t)$, respectivement.

Dans (8), les diverses grandeurs sont faciles à interpréter : $L_{y^{j,s}}$ et $U_{y^{j,s}}$ représentent les voisins inférieurs et supérieurs de t dans les valeurs échantillonnées $y^k, k \in s$, et $\beta_{y^{j,s}}(t)$ représente le coefficient d'interpolation linéaire entre ces deux grandeurs. En particulier, pour tout $t \in \{y^k, k \in s\}$, nous avons $H_{y^{j,s}}(t, y^k) = H(t - y^k)$. Par conséquent, les relations $F_{y^{j,s}}(t) = F_{y^{\text{cal}}}(t)$ sont satisfaites pour tout $t \in \{y^k, k \in s\}$. Pour toutes les autres valeurs de t , $F_{y^{\text{cal}}}(t)$ consiste en une interpolation linéaire entre ces grandeurs. Dans l'exemple qui suit, nous réexaminons l'exemple 1 en utilisant l'estimateur interpolé de la fonction de répartition (7).

Exemple 2

Dans l'exemple 1, si nous utilisons la version interpolée (7), les contraintes deviennent $w_1 + w_2 + w_3 = 30$ et $(w_1 + w_2) / (w_1 + w_2 + w_3) = 0,5$. Par conséquent, $w_3 = 15$, $w_1 + w_2 = 15$. Des opérations algébriques simples montrent que la solution optimale est $(w_1, w_2, w_3) = (10, 5, 15)$, qui est maintenant bien définie.

Si nous utilisons les estimateurs interpolés de la fonction de répartition, $F_{y^{j,s}}^{-1}(\alpha)$ et $F_{x^{j,s}}^{-1}(\alpha)$ sont maintenant des estimateurs bien définis du quantile α pour tout $\alpha \in (0, 1)$, à condition de pouvoir s'assurer que les poids w_k sont tous strictement positifs. En posant $\hat{Q}_{x^{j,s}, \text{cal}}^{\alpha} = F_{x^{j,s}, \text{cal}}^{-1}(\alpha)$, nous définissons l'estimateur par calage proposé $\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}$ pour le quantile $\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}$, en utilisant l'estimateur interpolé de la fonction de répartition donnée dans la définition 2.

Définition 3 (Estimateur par calage des quantiles). *Considérons le problème d'optimisation (3), sous les contraintes de calage $\sum_s y^k = N$ et $\hat{Q}_{x^{j,s}, \text{cal}}^{\alpha} = (\hat{Q}_{x^{j,s}, \text{cal}}^{\alpha})' = \mathbf{Q}_{x^{j,s}, \text{cal}}^{\alpha}$. Si nous résolvons ce problème d'optimisation et dénotons les poids résultant par w , l'estimateur par calage proposé des quantiles $\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}$ est défini par*

$$(9) \quad \hat{Q}_{y^{j,s}, \text{cal}}^{\alpha} = F_{y^{j,s}, \text{cal}}^{-1}(\alpha),$$

où $F_{y^{j,s}, \text{cal}}^{\alpha}(t)$ est donné par (6).

L'une des propriétés séduisantes de l'estimateur proposé (9) est qu'il donne des quantités de population exacts quand la relation entre y et une variable auxiliaire scalaire x est exactement linéaire. Émettons l'hypothèse que $y^k = a + bx^k$ tient parfaitement pour toutes les unités $k \in U$ et supposons que les unités de l'échantillon s sont telles que $x_k^j < \hat{Q}_{x^{j,s}, \text{cal}}^{\alpha} < x_l^j$ et $x_k^j, l^j \in s$. Pour nous devons faire la distinction entre les deux cas $b > 0$ et $b < 0$ (le cas $b = 0$ est trivial puisque y^k est alors identiquement égal à une constante). En premier lieu, considérons la situation $b > 0$. Comme la relation linéaire $y^k = a + bx^k$ est satisfait pour toutes les unités k et que $b > 0$, les relations qui suivent sont vérifiées : $L_{y^{j,s}}(a + bt) = a + bL_{x^{j,s}}(t)$; $U_{y^{j,s}}(a + bt) = a + bU_{x^{j,s}}(t)$ et $\beta_{y^{j,s}}(a + bt, y^k) = \beta_{x^{j,s}}(t)$. Ces relations mènent à $H_{y^{j,s}}(a + bt, y^k) = H_{x^{j,s}}(t, x^k)$. Il s'ensuit que $F_{y^{j,s}}(a + bt) = F_{x^{j,s}}(t)$. En outre, $F_{y^{j,s}, \text{cal}}^{\alpha}(a + b\hat{Q}_{x^{j,s}, \text{cal}}^{\alpha}) = \alpha$ et, en utilisant la relation $a + b\hat{Q}_{x^{j,s}, \text{cal}}^{\alpha} = \hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}$, nous déduisons que $F_{y^{j,s}, \text{cal}}^{\alpha}(\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}) = \alpha$. Par conséquent, lorsque une relation exactement linéaire est vérifiée et que $b > 0$, $\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha} = F_{y^{j,s}, \text{cal}}^{-1}(\alpha)$. En deuxième lieu, considérons le cas $b < 0$. Nous déduisons, dans ce cas, les relations qui suivent : $L_{y^{j,s}}(a + bt) = a + bL_{x^{j,s}}(t)$; $U_{y^{j,s}}(a + bt) = a + bU_{x^{j,s}}(t)$ et $\beta_{y^{j,s}}(a + bt, y^k) = 1 - \beta_{x^{j,s}}(t)$ et $H_{y^{j,s}}(a + bt, y^k) = 1 - H_{x^{j,s}}(t, x^k)$. Puisque $b < 0$, la relation entre les quantiles de x et de y est donnée par $a + b\hat{Q}_{x^{j,s}, \text{cal}}^{\alpha} = \hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}$. Alors, nous déduisons que $F_{y^{j,s}, \text{cal}}^{\alpha}(\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}) = F_{y^{j,s}, \text{cal}}^{\alpha}(a + b\hat{Q}_{x^{j,s}, \text{cal}}^{\alpha}) = 1 - \alpha$. Donc, dans cette situation, $\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}$ est estimée exactement par $\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}$. Autrement dit, lorsque une relation exacte est vérifiée, si $b > 0$, l'estimateur par calage proposé $\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}$ produit des estimateurs parfaits, de biais nul et de variance nulle. Par ailleurs, si $b < 0$ et que le calage est fait sur $\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}$, $\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}$ est estimée exactement par $\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}$ (ce qui est sensé, parce que la relation parfaitement linéaire entre x et y est telle que le paramètre de pente est négatif).

Notons que, si $F_{y^{j,s}, \text{cal}}^{\alpha}$ et $F_{x^{j,s}, \text{cal}}^{\alpha}$ sont interchangeables aux points $\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}$ et $\hat{Q}_{x^{j,s}, \text{cal}}^{\alpha}$, les contraintes de calage exprimées en (3) peuvent être réécrites par rapport aux fonctions de répartition, ce qui signifie que les contraintes de calage fondées sur les quantiles sont équivalentes à $F_{y^{j,s}, \text{cal}}^{\alpha}(\hat{Q}_{y^{j,s}, \text{cal}}^{\alpha}) = F_{x^{j,s}, \text{cal}}^{\alpha}(\hat{Q}_{x^{j,s}, \text{cal}}^{\alpha})$. Autrement dit, le problème de calage original peut être réexprimé par rapport aux fonctions de répartition avec les contraintes susmentionnées.

Une question naturelle est celle de savoir s'il existe une solution au problème d'optimisation (3). Même si celui-ci est formulé au moyen des fonctions de répartition interpolées, il n'est pas toujours possible de trouver une solution. Par exemple, si $\hat{Q}_{x^{j,s}, \text{cal}}^{\alpha}$ est plus petit ou plus grand

$$H(z) = \begin{cases} 1, & z \geq 0, \\ 0, & z < 0. \end{cases}$$

Nous définissons la fonction de répartition d'une variable auxiliaire scalaire x dans la population de la manière habituelle par $F_x^*(t) = N^{-1} \sum_{j=1}^J H(t - x_j^*)$, et nous obtenons la quantile de population $\tilde{Q}_{x,\alpha}$ en posant $\tilde{Q}_{x,\alpha} = \inf\{t | F_x^*(t) \geq \alpha\}$.

Le vecteur $\mathbf{Q}_{x,\alpha}$ contient les quantiles des variables auxiliaires, obtenus d'après l'information tirée d'enquêtes antérieures ou de sources administratives disponibles. Par exemple, pour les lois asymétriques qui sont assez fréquentes dans le cas des enquêtes auprès des entreprises et des enquêtes économiques, il paraît plus logique de garder dans les fichiers d'enregistrements les médianes plutôt que les moyennes de population; le cas échéant, il semble naturel de supposer que $\mathbf{Q}_{x,\alpha}$ est connu. Cela donne à penser qu'en suivant la même approche que celle menant au calage des totaux décrits à la section 2, l'estimateur proposé pour les quantiles de population $\tilde{Q}_{y,\alpha}$ de la variable d'intérêt y , noté $\tilde{Q}_{y,\alpha}$, pourrait être obtenu par inversion d'un certain estimateur de la fonction de répartition (dont nous discutons plus loin), sous des contraintes de calage telles que $\tilde{Q}_{x_j,\alpha} = \tilde{Q}_{y_j,\alpha}$, $j = 1, \dots, J$. Suivant l'interprétation habituelle, si les poids calés nous permettent d'extraire les quantiles de population connus des variables auxiliaires, alors, sous certaines conditions, ils devraient produire des estimateurs raisonnables des quantiles de la variable d'intérêt y .

Plus précisément, nous obtenons les poids calés en résolvant le problème d'optimisation suivant :

$$(3) \quad \mathbf{w} = \arg \min_y D(\mathbf{y}, \mathbf{d}),$$

sous les contraintes de calage $\sum_{j=1}^J w_j = N$ et $\tilde{Q}_{x,\alpha} = \tilde{Q}_{y,\alpha}$, \dots , $\tilde{Q}_{x_J,\alpha} = \tilde{Q}_{y_J,\alpha}$. Les estimateurs $\tilde{Q}_{x_j,\alpha}$ et $\tilde{Q}_{y_j,\alpha}$ s'appuient sur le vecteur de poids \mathbf{w} , issus de la résolution du problème de calage (3). Pour calculer ces estimateurs pour les quantiles, nous devons construire les estimateurs pondérés par les poids \mathbf{w} de la fonction de répartition des variables \mathbf{x} et y . En se basant sur les poids d'échantillonnage \mathbf{d} , un estimateur logique de la fonction de répartition d'échantillonnage est donné par

$$F_y^*(t) = \sum_{j=1}^J d_j H(t - y_j^*) / \sum_{j=1}^J d_j, \quad (4)$$

qui fournit un estimateur convergent de $F_y(t)$. De même, $F_x^*(t) = \sum_{j=1}^J d_j H(t - x_j^*) / \sum_{j=1}^J d_j$, $j = 1, \dots, J$. Un estimateur de la fonction de répartition pondéré par les poids \mathbf{w} de $F_{y_j}^*(t)$ est donné par

$$F_{y_j,\text{cal}}^*(t) = \sum_{j=1}^J w_j H(t - x_j^*) / \sum_{j=1}^J w_j. \quad (5)$$

Une formule similaire est vérifiée pour $F_{y,\text{cal}}^*(t)$. Ces estimateurs pondérés par les poids \mathbf{w} sont considérés dans Ren (2002). Cependant, si l'on estime $\tilde{Q}_{x_j,\alpha}$ par $\tilde{Q}_{x_j,\alpha} = \inf\{t | F_{x_j}^*(t) \geq \alpha\}$, on que l'on fait une estimation similaire en utilisant une version pondérée par les poids \mathbf{w} , alors il est généralement impossible d'atteindre une solution exacte du problème de calage (3). En effet, si l'on utilise la définition qui précède pour estimer les quantiles par inversion de la fonction de distribution en utilisant les définitions antérieures, les contraintes du problème d'optimisation (3) ne seront généralement pas remplies, à moins que l'échantillon s contienne précisément une unité k telle que $x_k^* = \tilde{Q}_{x_j,\alpha}$. Quand J est grand, ce problème peut être plus prononcé. En outre, même si l'échantillon contient une telle valeur, il est parfois impossible d'obtenir les poids nécessaires pour minimiser la fonction de distance, parce que, dans certaines circonstances, les poids qui satisfait les contraintes de calage forment un ensemble ouvert, tandis que les poids optimaux se situent précisément sur la limite de cet ensemble. L'exemple qui suit illustre cette situation.

Exemple 1

Considérons une population U de taille $N = 30$, telle que la médiane de population de x soit $\tilde{Q}_{x,0.5} = 2$. Nous tirons un échantillon s de taille $n = 3$ et supposons que $x_k = k$, $\forall k \in s = \{1, 2, 3\}$. Pour simplifier, nous adoptons la mesure de distance $D(\mathbf{y}, \mathbf{d}) = \sum_{j=1}^J (y_j - d_j^*)^2$, nous la contraindre de (5), la contrainte de calage est $\tilde{Q}_{x,\text{cal},0.5} = \inf\{t | F_{x,\text{cal}}^*(t) \geq 0.5\} = 2$, qui implique $\sum_{j=1}^J w_j H(2 - x_j^*) \geq 15$ et $\sum_{j=1}^J w_j H(1 - x_j^*) < 15$. De façon équivalente, $w_1 + w_2 \geq 15$ et $w_1 < 15$. Donc, nous devons choisir w_1 de la forme $w_1 = 15 - \epsilon$, pour $\epsilon > 0$. Dans ce cas, puisque $w_1 + w_2 + w_3 = 30$, nous avons que $D(\mathbf{y}, \mathbf{d}) = \epsilon^2 + (w_2 - 9)^2 + (w_2 - 9 - \epsilon)^2$, ce qui nous mène à la solution optimale $(w_1, w_2, w_3) = (15 - \epsilon, 9 + \epsilon/2, 6 + \epsilon/2)$. Par conséquent, pour ces poids, $D(\mathbf{y}, \mathbf{d}) = 3\epsilon^2/2$, qui est de toute évidence minimisé quand $\epsilon \rightarrow 0$. Toutefois, la limite se réduit à $\mathbf{w} = (w_1, w_2, w_3) = (15, 9, 6)$, avec $D(\mathbf{w}, \mathbf{d}) = 0$, mais, d'après ces poids, $\tilde{Q}_{x,\text{cal},0.5} = 1 \neq \tilde{Q}_{x,0.5} = 2$.

Néanmoins, il est possible d'éviter ces difficultés en envisageant un estimateur lisse de la fonction de répartition. Pour simplifier, considérons ici un estimateur de la fonction de répartition calculé par interpolation linéaire (nous discutons d'une autre possibilité à la section 5), qui est défini précisément dans la définition 2.

Définition 2 (Estimateur interpolé de la fonction de répartition). *Définitions*

$$F_{y,\text{cal}}^*(t) = \frac{\sum_{j=1}^J w_j H_{y_j,s}(t, y_j^*)}{\sum_{j=1}^J w_j}, \quad (6)$$

courant, en pratique, de poser $x_k \equiv 1, \forall k \in U$, et par conséquent $T_x = N$. Cela signifie que les poids calés satisfont la contrainte naturelle $\sum_s w_k = N$. De nombreuses fonctions de distance D sont proposées dans la littérature (voir, par exemple, Deville et Särndal (1992); Chen et Qin (1993); Thompson (1997)). Considérons la fonction de distance quadratique.

$$(2) \quad D(v, p) = \sum_s \frac{(v_s^k - p_s^k)^2}{p_s^k q_s^k},$$

où q_k détermine l'importance de l'unité $k \in s$ dans le problème de calage. Les problèmes d'hétéroscédasticité peuvent être réglés en choisissant les valeurs de q_k de façon appropriée. En résolvant le problème d'optimisation (1) par la technique du multiplicateur de Lagrange (voir Deville et Särndal 1992, entre autres), nous obtenons les poids $w_k = d_k^k (1 + g_k^k x_k^k \lambda_s)$, où $\lambda_s = (\sum_s d_k^k q_k^k x_k^k)^{-1} (T_x - T_{x,HT})$ et $T_{x,HT}$ représente l'estimateur HT de T_x . Ce choix de la fonction de distance aboutit aux poids de l'estimateur par la régression généralisée (GREG) bien connu de Cassel, Särndal et Wretman (1976), qui est étudiée en détail dans Särndal, Swensson et Wretman (1992). Sous des exigences minimales concernant la mesure de distance D , Deville et Särndal (1992) ont montré que tous les estimateurs par calage de cette classe sont asymptotiquement équivalents à l'estimateur GREG. Pour faciliter l'interprétation et pour d'autres raisons esthétiques, certains utilisateurs pourraient souhaiter obtenir des poids positifs ou les contraindre à un intervalle particulier (voir aussi Singh et Mohl 1996). Dans les applications pratiques, ces caractéristiques numériques des poids semblent être le motif principal de divers choix de D .

3. Nouveaux estimateurs par calage

À la présente section, nous élaborons des estimateurs par calage pour les quantités, selon des idées similaires à celles dominant lieu aux estimateurs par calage des totaux de population décrits à la section 2. Nous présentons les nouveaux estimateurs par calage pour les quantités à la sous-section suivante, en utilisant des estimateurs interpolés de la fonction de répartition. Puis, nous accordons une attention spéciale à la fonction de distance quadratique. À la dernière sous-section, nous présentons l'estimation de la variance et la construction des intervalles de confiance.

3.1 Définition des estimateurs par calage des quantités

Soit $Q_{x,a} = (Q_{x,a}^1, \dots, Q_{x,a}^J)'$ le vecteur connu des quantités de population pour le vecteur de variables auxiliaires $x_k = (x_{1k}, \dots, x_{Jk})'$, $k \in U$. La fonction de Heaviside $H(z)$ est donnée par :

Chambers et Dunstan (1986), ainsi qu'à certains estimateurs proposés par Rao et coll. (1990). Enfin, à la section 5, nous présentons nos conclusions.

2. Certains préliminaires sur les estimateurs par calage

À la présente section, nous présentons les concepts fondamentaux et les notations qui seront utiles dans la suite. En outre, nous passons brièvement en revue les estimateurs par calage des totaux.

Soit $U = \{1, \dots, k, \dots, N\}$ une population finie de taille N . Soit $T_y = \sum_U y_k$ le total de population de la variable d'intérêt y (notons que pour un ensemble A , $A \subseteq U$, nous utiliserons \sum_A comme abréviation de $\sum_{k \in A}$). Nous tirons un échantillon $s \subset U$ de taille n conformément à un plan d'échantillonnage. Soit $\pi_k = \Pr(s \ni k)$ et $\pi_s = \Pr(s \ni k, \dots, k_s)$ les probabilités d'inclusion de premier et de deuxième ordre, respectivement. Nous dénotons les poids de sondage par $d_k^k = \pi_k^{-1}$ et $T_{y,HT} = \sum_s d_k^k y_k$ représente l'estimateur d'Horvitz-Thompson (HT) de T_y .

Soit $x_k = (x_{1k}, \dots, x_{Jk})'$ un vecteur de variables auxiliaires associé à l'unité k , $k \in U$. Les estimateurs par calage incluent naturellement l'information auxiliaire dans l'estimation. Soit $s = \{k_1, \dots, k_n\}$, $s \subset U$. En partant du vecteur de poids originaux $p = (d_{k_1}^k, \dots, d_{k_n}^k)'$, nous trouvons de nouveaux poids qui, lorsqu'ils sont appliqués aux variables auxiliaires disponibles dans s , permettent d'extraire les totaux de population connus pour les J variables auxiliaires $T_x = \sum_U x_k = (T_{x_1}, \dots, T_{x_J})'$. Les estimateurs par calage des totaux seront définis de façon plus précise dans la définition 1.

Définition 1 (estimateur par calage des totaux). Soit $d = (d_{k_1}, \dots, d_{k_n})'$ les poids de sondage. L'estimateur par calage des totaux prend la forme $T_{y,cal} = \sum_s w_k y_k$, où les poids w_k , $k \in s$ sont obtenus en résolvant le problème de minimisation qui suit par rapport à la variable $v = (v_{k_1}, \dots, v_{k_n})'$:

$$(1) \quad w = \arg \min_v D(v, d),$$

sous les contraintes de calage $\sum_s w_k x_k = T_x$, où $D(\cdot, \cdot)$ représente la mesure de distance et $w = (w_{k_1}, \dots, w_{k_n})'$ correspond au vecteur des poids calés.

Pour simplifier la notation, nous écrivons $w_k \equiv w_{k_s}$ dans la définition 1 quand aucune confusion n'est possible. Il est

que l'on suppose connus. Nous présentons des arguments qui justifient le calage sur les quantiles, quand le paramètre d'intérêt est lui-même un quantile. Fait intéressant, notre méthode ne nécessite pas que l'on connaisse les valeurs des variables auxiliaires pour toutes les unités de la population. Puisque les estimateurs résultants présentent une forme structurelle fort semblable à celle des estimateurs par calage originaux des totaux, nous nous attendons à ce que, sous des conditions générales, les estimateurs proposés des quantiles soient asymptotiquement sans biais par rapport au plan de sondage. En outre, ces similarités nous permettent de dériver des estimateurs de la variance qui admettent une forme familière. Contrairement à certains autres estimateurs, l'approche proposée est également applicable aux variables auxiliaires vectorielles (c'est-à-dire les situations où plusieurs variables auxiliaires sont disponibles), tout en ne requérant que des renseignements auxiliaires minimes. Cependant, certaines restrictions pourraient s'appliquer lorsque l'échantillon est fortement non représentatif de la population ou que les quantiles que l'on estime sont très proches du minimum ou du maximum de population. Notons que des échantillons fortement non représentatifs peuvent aussi causer des problèmes dans le cas des estimateurs par calage des totaux utilisés couramment; le cas échéant, l'algorithme pour le calcul de ces estimateurs peut ne pas converger pour de nombreuses mesures de distance présentant un intérêt pratique.

La présentation de l'article est la suivante. À la section 2, nous donnons certains préliminaires, dont un bref examen des estimateurs par calage des totaux. À la section 3.1, nous décrivons l'élaboration des nouveaux estimateurs par calage des quantiles. La fonction de répartition standard peut être interprétée comme étant un estimateur d'Horvitz-Thompson, qui offre une approche possible de la construction d'un estimateur calé de la fonction de répartition. Les estimateurs des quantiles sont alors dérivés naturellement par inversion de l'estimateur de la fonction de répartition (voir, par exemple, Ren (2002)). Comme dans le cas des estimateurs par calage des totaux, les poids de sondage peuvent être remplacés par des poids d'échantillonnage plus généraux, afin de tenir compte de l'information auxiliaire. Toutefois, dans de nombreuses situations présentant un intérêt pratique, il se peut qu'aucune solution n'existe pour les contraintes de calage lorsqu'on adopte ce genre d'estimateur des fonctions de répartition, parce que celui-ci correspond à une fonction échelon. Afin d'éviter les problèmes d'existence de solutions pour les contraintes de calage, nous présentons un nouvel estimateur de la fonction de répartition fondé sur le concept naturel d'interpolation. À la section 3.2, nous présentons, sous les conditions de la métrique quadratique ordinaire, une représentation analytique des poids de calage; à la section 3.3, nous discutons des estimateurs de la

est devenue populaire dans les applications pratiques, parce que les estimateurs résultants sont faciles à interpréter et à justifier, étant donné qu'ils s'appuient sur les poids d'échantillonnage et des contraintes de calage naturelles. Cette approche a été élaborée dans le cadre des travaux fondateurs de Deville et Särndal (1992) à titre de nouveau moyen d'intégrer l'information auxiliaire dans l'estimation des totaux de population. Les poids dits calés sont obtenus en minimisant une mesure de distance entre les poids d'échantillonnage et les nouveaux poids sous certaines contraintes de calage. Pour l'estimation des totaux, les poids calés remplacent les poids de sondage originaux utilisés dans les estimateurs de type Horvitz-Thompson. Lorsqu'on les applique aux variables auxiliaires disponibles dans l'échantillon, les nouveaux poids reproduisent exactement les totaux connus de population de ces variables, d'où le nom d'estimateurs par calage donné aux estimateurs de cette classe. Voir aussi Singh et Mohl (1996) qui fournissent des justifications simples des estimateurs par calage. Ils présentent en outre un traitement très général et harmonisé des méthodes par calage produisant des poids qui satisfont à certaines restrictions concernant les fourchettes de valeurs et certaines contraintes d'égalonage.

Essentiellement, notre but est de proposer pour les quantiles des estimateurs par calage aussi faciles à appliquer et à interpréter que les estimateurs par calage des totaux mis au point par Deville et Särndal (1992). Comparativement aux estimateurs des quantiles décrits dans la littérature, les nouveaux estimateurs par calage devraient aussi donner des résultats avantageux en ce qui concerne le biais, la variance et les taux de couverture des intervalles de confiance. Les premiers estimateurs par calage proposés pour les fonctions de répartition et les quantiles comprennent ceux de Kwončević (1997), qui a étudié des estimateurs de la fonction de répartition calés sur les moments des variables auxiliaires; Harms (2003) a suivi une approche analogue comportant des applications à la version finlandaise du Panel européen des ménages. Ren (2002) semble avoir été le premier à élaborer un traitement unifié de l'application des estimateurs par calage aux fonctions de répartition et aux quantiles. Les estimateurs par calage applicables aux quantiles présentés ici constituent une prolongation des travaux de Ren (2002). Nous adhérons aussi étroitement que possible au paradigme de calage original appliqué aux totaux : lorsque le paramètre d'intérêt est un total, il semble logique de faire le calage sur les variables auxiliaires. Ici, puisque le paramètre d'intérêt est un quantile, les contraintes de calage imposent l'utilisation de poids tels que les estimateurs des quantiles d'échantillon des variables auxiliaires et de leurs quantiles de population correspondants soient égaux. Autrement dit, les estimateurs pondérés des quantiles des variables auxiliaires devraient produire exactement les quantiles de population,

De l'estimation des quantiles par calage

Torsten Harms et Pierre Duchesne

Résumé

Le présent article traite de l'application du paradigme de calage à l'estimation des quantiles. La méthodologie proposée suit une approche semblable à celle qui donne lieu aux estimateurs par calage originaux de Deville et Sarda (1992). Une propriété intéressante de cette nouvelle méthodologie est qu'elle ne nécessite pas la connaissance des valeurs des variables auxiliaires pour toutes les unités de la population. Il suffit de connaître les quantiles correspondants de ces variables auxiliaires. L'adoption d'une métrique quadratique permet d'obtenir une représentation analytique des poids de calage, qui sont alors similaires à ceux menant à l'estimateur par la régression généralisée (GREG). Nous discutons de l'estimation de la variance et de la construction des intervalles de confiance. Au moyen d'une petite étude par simulation, nous comparons l'estimateur par calage à d'autres estimateurs fréquemment utilisés des quantiles qui s'appuient également sur des données auxiliaires.

Mots clés : Estimateurs par calage; estimateurs par la différence; estimateurs par le ratio; quantiles.

1. Introduction

Ces dernières années, dans le contexte des sondages, beaucoup d'attention a été accordée à l'estimation des fonctions de répartition des populations. À cet égard, la médiane, souvent considérée comme une mesure d'emplacemement plus satisfaisante que la moyenne, particulièrement si la variable d'intérêt suit une loi asymétrique, a suscité un intérêt particulier. Habituellement, les estimateurs traditionnels des moyennes et des totaux de population peuvent être améliorés sensiblement si l'on dispose d'information auxiliaire pertinente. Par conséquent, il paraît fort souhaitable d'utiliser ce genre d'information dans les estimateurs

Adoptant une approche basée sur un modèle, Chambers et Dunstan (1986) ont examiné des estimateurs des quantiles et Dunstan (1986) ont examiné des estimateurs des quantiles basés sur un estimateur de la fonction de répartition auquel sont intégrés des renseignements auxiliaires. Rao, Kovar et Mantel (1990) ont proposé des variantes fondées sur le plan de sondage de l'approche basée sur un modèle. Ils ont comparé, dans des expériences par simulation, deux estimateurs des quantiles, basés sur des estimateurs par le ratio et par différence, à l'estimateur fondé sur le plan de sondage simple dans lequel n'est utilisée aucune information auxiliaire. Il convient de souligner que ni l'un ni l'autre des estimateurs proposés fondés sur le plan de sondage ne nécessite la connaissance de l'information auxiliaire pour chaque unité de la population; il suffit de connaître les quantiles correspondants. Bien que l'estimateur basé sur un modèle proposé par Chambers et Dunstan (1986) puisse être plus efficace que l'option fondée sur le plan de sondage si le modèle est spécifié correctement, Rao et coll. (1990) soulignent l'avantage des estimateurs fondés sur le plan de

sondage en cas de spécification incorrecte du modèle. Chambers, Dorfman et Hall (1992) ont comparé théoriquement la convergence, le biais asymptotique et la variance de ces deux estimateurs sous un modèle de population. Leur conclusion principale est qu'aucune des deux méthodes n'est sensiblement meilleure que l'autre. Dorfman (1993) a réévalué les résultats des simulations de Rao et coll. (1990) et proposé une version modifiée de leur méthode reposant sur des arguments fondés sur un modèle. Les estimateurs de la variance dans le contexte de l'approche basée sur un modèle de Chambers et Dunstan (1986) et des estimateurs fondés sur le plan de sondage de Rao et coll. (1990) sont examinés dans Wu et Sitter (2001).

Parmi les autres travaux relatifs aux estimateurs des quantiles et de la médiane, mentionnons ceux de Kuk (1988) qui propose des estimateurs des quantiles sous échantillonnage PPT (*probability proportionnelle à la taille*) et ceux de Kuk et Mak (1989) qui utilisent une méthode basée sur la classification croisée des individus compris dans l'échantillon en fonction de la variable d'intérêt et d'une variable auxiliaire unique. Meeeden (1995) adopte une approche différente pour construire un estimateur de la médiane basé sur des données auxiliaires univariées, en utilisant le concept bayésien de l'échantillonnage de Pólya pour imputer toutes les valeurs de population inconnues de la variable cible selon une approche fondée sur le ratio. Récemment, Rueda, Arcos et Martínez (2003) ont construit des estimateurs des quantiles qui étendent les estimateurs par le ratio, par la différence et par la régression de façon similaire à ceux élaborés pour la moyenne de population. Dans le présent article, nous appliquons le concept de calage que Deville (1988) a été le premier à proposer afin de dériver un estimateur des quantiles. L'approche par calage

- Luo, M., Stokes, L. et Sager, T. (1998). Estimation of the CDF of a finite population in the presence of a calibration sample. *Environmental and Ecological Statistics*, 5, 277-289.
- Moore, J.C., Stinson, L.L. et Welniak, E.J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics*, 16, 331-361.
- ONS (1999). *Labour Force Survey. User Guide, Volume 1, Background and Methodology*. London.
- Ramcourt, E. (1999). Estimation with nearest neighbour imputation at Statistics Canada. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 131-138.
- Rodgers, W.L., Brown, C. et Duncan, G.J. (1993). Errors in survey reports of earnings, hours worked and hourly wages. *Journal of the American Statistical Association*, 88, 1208-1218.
- Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.
- Rubin, D.B., et Schenker N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable non-response. *Journal of the American Statistical Association*, 81, 366-374.
- Skinner, C.J. (1989). Domain means, regression and multivariate analysis. Dans *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt et T.M.F. Smith), Chichester, Wiley.
- Skinner, C., Stutard, N., Beissel-Durrant, G. et Jenkins, J. (2002). The measurement of low pay in the UK Labour Force Survey. *Oxford Bulletin of Economics and Statistics*, 64, 653-676.
- Stutard, N., et Jenkins, J. (2001). Measuring low pay using the new earnings survey and the Labour Force Survey. *Labour Market Trends*, janvier 2001, 55-66.
- Tenenbein, A. (1970). A double sampling scheme for estimating from binary data with misclassifications. *Journal of the American Statistical Association*, 65, 1350-1361.
- David, M.H., Little, R., Samuël, M. et Triest, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 168-173.
- Dickens, R., et Manning, A. (2004). Has the national minimum wage reduced UK wage inequality? *Journal of the Royal Statistical Society, Series A*, 4, 613-626.
- Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.
- Fay, R.E. (1999). Theory and application of nearest neighbour imputation in census 2000. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 112-121.
- Fuller, W.A. (1995). Estimation in the presence of measurement error. *Revue Internationale de Statistique*, 63, 121-141.
- Kalton, G. (1983). *Compensating for missing survey data*. Michigan, Institute for Social Research.
- Kalton, G., et Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics, Part A, Theory and Methods*, 13, 1919-1939.
- Kim, J.K. (2004). Efficient nonresponse weighting adjustment using estimated response probability. *Proceedings of the Survey Research Methods Section, American Statistical Association*.
- Kim, J.-K., et Fuller, W.A. (2002). Variance estimation for nearest neighbour imputation. Manuscript non-publié.
- Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 139-157.
- Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-301.
- Little, R.J.A., et Rubin, D.B. (2002). *Statistical analysis with missing data*. New York: John Wiley & Sons, Inc.

Tableau 6 Estimations de θ_1 et θ_2 (pondérées) pour le groupe des 18 ans et plus par application de divers modèles de pondération par le score de propension et d'imputation aux données de l'EPA, juin à août 1999

Méthode	Modèle de pondération par le score de propension	(Pondérée)	(Pondérée)	θ_1 (%)	θ_2 (%)
Variable dérivée	—	7,13	20,5		
Pondération par le score de propension	M1	0,96	34,5		
	M2	1,08	38,4		
	M3	1,08	38,4		
IHDR10	M1	1,44	32,1		
	M2	1,41	32,9		
	M3	1,50	33,2		
PPV10	M1	1,32	32,6		
	M2	1,44	32,8		
	M3	1,50	33,0		

Nota : M1 est le modèle le plus complexe comprenant des termes quadratiques et des termes d'interactions. M2 exclut les termes d'interactions et les termes quadratiques inclus dans le modèle M1. M3 correspond à l'élimination de covariables supplémentaires par rapport au modèle M2.

Tableau 7 Estimations de θ_1 et θ_2 (pondérées) pour le groupe des 18 ans et plus par application de la pondération par le score de propension et de l'imputation aux données de l'EPA, mars à mai 2000

Méthode	Modèle de pondération par le score de propension ou modèle d'imputation	(Pondérée)	(Pondérée)	θ_1 (%)	θ_2 (%)
Pondération par le score de propension	M1	0,54	27,10		
IHDR10	M1	0,57	26,01		
PPV10	M1	0,55	26,61		

Nous avons entrepris d'autres travaux en vue d'élaborer et d'évaluer des méthodes d'estimation de la variance associées, ainsi que d'autres méthodes d'estimation ponctuelle fondées sur le modèle commun d'erreur de mesure décrit à la section 2.

Remerciements

Nous remercions Danny Pfeffermann pour ses commentaires concernant une version antérieure du présent article.

Bibliographie

Brick, J.M., et Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Buonaccorsi, J.P. (1990). Double sampling for exact values in some multivariate measurement error problems. *Journal of the American Statistical Association*, 85, 1075-1082.

Chen, J., et Shao, J. (2000). Nearest neighbour imputation for survey data. *Journal of Official Statistics*, 16, 113-131.

Chen, J., et Shao, J. (2001). Jackknife variance estimation for nearest neighbour imputation. *Journal of the American Statistical Association*, 96, 453, 260-269.

Chesher, A. (1991). The effect of measurement error. *Biometrika*, 78, 451-462.

Dans le présent article, nous avons examiné l'application de diverses méthodes de traitement des données manquantes en vue de corriger le biais causé par l'erreur de mesure dans l'estimation d'une fonction de distribution. Parmi les méthodes d'imputation, celles par le plus proche voisin donnent les résultats les plus prometteurs en ce qui concerne le biais. Il n'existe aucun signe que ces méthodes déterministes produisent un biais plus important que les méthodes d'imputation stochastiques. L'imputation fractionnaire donne lieu à des gains d'efficacité appréciables comparativement à l'imputation simple et semble être plus efficace que la pénalisation de la fonction de distance ou que l'échantillonnage sans remise avec imputation simple. Comparativement à la méthode de pondération par le score de propension, l'imputation fractionnaire par le plus proche voisin donne des résultats comparables, mais présente de légers avantages en ce qui a trait à la robustesse et à l'efficacité. L'étude en simulation laisse entendre que l'effet sur le biais sous un modèle mal spécifié est plus important dans le cas de la pondération par le score de propension et que les erreurs-types sous l'approche de pondération sont supérieures de 5 % à 15 % à celles observées pour la méthode d'imputation.

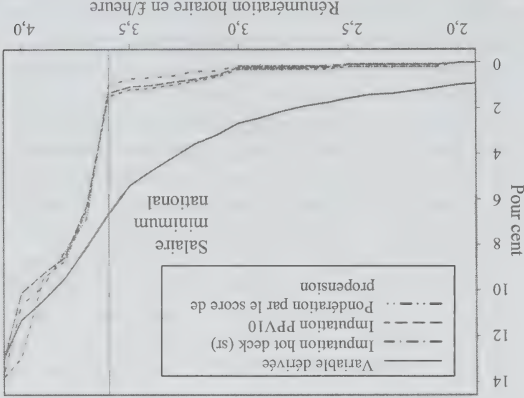
9. Conclusion

Les méthodes décrites aux sections 2 à 4 ont été élaborées sous l'hypothèse d'un modèle IID et d'un échantillonnage ignorable. Dans le cas de l'EPA, les employés sont sélectionnés avec probabilités égales, de sorte que l'échantillonnage peut être considéré comme ignorable en ce qui concerne le biais de l'estimation ponctuelle, mais la non-réponse totale est vraisemblablement différentielle, de sorte que les poids de sondage sont calculés de façon à en tenir compte (ONS 1999). Nous proposons d'intégrer ces poids de sondage dans l'estimateur (3) ou, de façon équivalente, de multiplier les coefficients de pondération w_i dans (9) par les poids de sondage. Cette approche est analogue à la façon dont les estimateurs sont pondérés en se fondant sur une hypothèse IID dans l'approche de la pseudosurressemblance (Skinner 1989). Le but est d'utiliser les méthodes décrites aux sections 2 à 4 pour corriger le biais dû à l'erreur de mesure et à la non-réponse partielle, ainsi que les poids de sondage pour corriger le biais dû à l'échantillonnage et à la non-réponse totale. Nous n'avons pas essayé de tenir compte des pondérations dans les méthodes d'imputation et cette question pourrait être étudiée dans le cadre de futurs travaux.

Nous appliquons maintenant l'imputation par le plus proche voisin, l'imputation hot deck dans les classes et la pondération par le score de propension aux données de l'EPA. Nous appliquons la pondération par les poids de sondage à toutes les méthodes. À la figure 1, nous comparons une distribution estimée, qui ne tient pas compte de l'erreur de mesure (courbe en trait plein) aux estimations obtenues par les trois méthodes de traitement des données manquantes (les trois courbes en trait interrompu). Nous soutenons que l'absence approximative de biais est plus importante pour ces dernières estimations que pour la première. Les trois ajustements pour les données manquantes, comme prévu, une forte « cassure » dans la distribution au niveau de la rémunération horaire minimale nationale correspondant à la variable dérivée. Les estimations correspondantes de deux proportions de faible rémunération horaire d'intérêt sont présentées au tableau 6. Les « ajustements pour les données manquantes » ont un effet appréciable comparativement aux estimations fondées sur la variable dérivée. Les résultats laissent entendre que la proportion d'emplois rémunérés au taux horaire minimal national ou à un taux inférieur pourrait être surestimée d'un facteur quatre ou cinq si l'on ne tient pas compte de l'erreur de mesure. Les écarts entre les méthodes de traitement des données manquantes sont nettement plus faibles. Nous voyons que les estimations sous la pondération par le score de propension diffèrent des estimations calculées par les méthodes d'imputation, du moins pour le trimestre de juin à août 1999. Il convient de souligner que le taux de réponse

durant ce trimestre de l'EPA a été plus faible que pour les trimestres suivants à cause de modifications apportées au questionnaire de l'enquête. Nous avons constaté que pour les trimestres suivants, la pondération et l'imputation ont donné des estimations fort semblables des proportions de travailleurs faiblement rémunérés, comme l'illustre le tableau 7 pour le trimestre de mars à mai 2000. La diminution de la proportion d'emplois faiblement rémunérés au cours du temps est due à l'effet de la loi sur le salaire minimum national. En outre, nous utilisons différents modèles d'imputation et de pondération par le score de propension en vue d'analyser les effets de diverses spécifications du modèle sur les estimations de la proportion d'emplois faiblement rémunérés. Le tableau 6 montre que l'utilisation de différents modèles pourrait avoir une incidence sur les estimations. À mesure que le modèle devient plus complexe, nous observons une réduction des estimations dans le cas des deux estimateurs ponctuels. Cela pourrait refléter un écart par rapport à l'hypothèse de mécanisme MAR pour les modèles d'imputation plus simples. Du moins pour le trimestre de 1999, les différences de méthodes de pondération et d'imputation qu'entre les modèles. Soulignons que les estimations présentées ici pourraient différer légèrement des estimations officielles pour le Royaume-Uni, puisque, par exemple, les estimations officielles sont fondées sur des modèles d'imputation différents, qui traitent différemment les valeurs extrêmes ou qui imputent différemment les valeurs pour certaines professions.

Figure 1. Diverses estimations de la distribution de la rémunération horaire de 2 £ à 4 £ pour le groupe des 22 ans et plus, juin à août 1999.



À la présente section, nous examinons l'application des méthodes élaborées aux sections 2 à 4 aux données de l'EPA. Cette dernière représente une source importante de données pour l'estimation de la rémunération horaire au Royaume-Uni (Stuttard et Jenkins 2001). Il s'agit d'une enquête trimestrielle réalisée auprès de ménages sélectionnés à partir d'un fichier national d'adresses postales avec probabilités égales par échantillonnage systématique stratifié. Tous les adultes compris dans les ménages sélectionnés sont inclus dans l'échantillon. L'échantillon résultant est mis en grappes selon l'appartenance au ménage, mais non selon les caractéristiques géographiques. Chaque ménage sélectionné demeure dans l'échantillon en vue d'être interviewé pendant cinq trimestres consécutifs, puis est éliminé de l'échantillon et remplacé. Les questions concernant la rémunération horaire sont posées durant les première et cinquième interviews seulement, ce qui produit des données à ce sujet pour environ 16 000 employés par trimestre.

Deux mesures de la rémunération horaire sont considérées, comme il est décrit à la section 1. La variable dérivée de la rémunération horaire de l'EPA est définie comme suit : a) des questions sont posées aux employés au sujet de leur emploi principal afin de déterminer les gains au cours d'une période de référence, b) des questions sont posées afin de déterminer le nombre d'heures travaillées au cours de la période de référence et c) le résultat de a) est divisé par le résultat de b). La variable directe est obtenue en commentant par demander si le répondant est payé à un taux horaire

8. Application à l'Enquête sur la population active

Méthode	Biais de θ_1	Biais rel. de θ_1	Biais de θ_2	Biais rel. de θ_2	Pondération par le score de propension
PPV10	$29,0 \times 10^{-4}$	5,1 %	$92,0 \times 10^{-4}$	5,0 %	$0,8 \times 10^{-4}$
	$(0,73 \times 10^{-4})$		$(1,48 \times 10^{-4})$		$(0,73 \times 10^{-4})$
			100×10^{-4}		
			$(1,40 \times 10^{-4})$		
				5,7 %	
				5,7 %	
					$2,31 \times 10^{-3}$
					$4,42 \times 10^{-3}$
					$e.l.(\theta_2)$
					$2,53 \times 10^{-3}$
					$e.l.(\theta_1)$
					$4,70 \times 10^{-3}$

Tableau 5
Estimations par simulation des biais et des erreurs-types des estimateurs de θ_1 et θ_2 pour l'imputation par le plus proche voisin (PPV10) et la pondération par le score de propension, sous le modèle (non-MAR) commun d'erreur de mesure ($H = 1\ 000$)

Méthode	Covariables hypothétiques	$e.l.(\theta_1)$	$e.l.(\theta_2)$	$EQM(\theta_1)$	$EQM(\theta_2)$
PPV10	M1 (correct)	$2,02 \times 10^{-3}$	$3,80 \times 10^{-3}$	$4,10 \times 10^{-6}$	$1,49 \times 10^{-5}$
	M2	$2,06 \times 10^{-3}$	$3,88 \times 10^{-3}$	$4,29 \times 10^{-6}$	$1,54 \times 10^{-5}$
	M3	$2,01 \times 10^{-3}$	$3,89 \times 10^{-3}$	$4,10 \times 10^{-6}$	$1,63 \times 10^{-5}$
	M1 (correct)	$2,27 \times 10^{-3}$	$4,27 \times 10^{-3}$	$5,16 \times 10^{-6}$	$1,83 \times 10^{-5}$
	M2	$2,17 \times 10^{-3}$	$4,42 \times 10^{-3}$	$5,51 \times 10^{-6}$	$6,90 \times 10^{-5}$
	M3	$2,16 \times 10^{-3}$	$4,46 \times 10^{-3}$	$4,94 \times 10^{-6}$	$6,59 \times 10^{-5}$

Tableau 4
Estimations par simulation des erreurs-types des estimateurs de θ_1 et θ_2 pour l'imputation par le plus proche voisin (PPV10) et la pondération par le score de propension, sous l'hypothèse de mécanisme MAR et de covariables correctes et spécifiées incorrectement ($H = 1\ 000$)

Les estimations correspondantes des erreurs-types de θ_1 et θ_2 sont données au tableau 4. Celles-ci ont aussi tendance à être plus importantes pour la méthode de pondération, l'accroissement étant de 5 % à 15 % par rapport à la méthode d'imputation. L'augmentation de l'erreur-type est plus importante pour le deuxième estimateur θ_2 , variant de 12 % à 15 %, que pour l'estimateur θ_1 , pour lequel l'accroissement est de 5 % à 12 %, selon l'erreur de spécification. Par conséquent, l'erreur quadratique moyenne est également plus grande pour la méthode de pondération, l'accroissement variant de 20 % à 28 % pour les six valeurs au tableau 4. Du moins sous l'hypothèse de mécanisme MAR, la méthode d'imputation PPV10 semble être préférable à la pondération par le score de propension en ce qui concerne le biais et la variance.

Enfin, nous comparons les propriétés des méthodes d'imputation (PPV10) et de pondération par le score de propension lorsque l'hypothèse de mécanisme MAR ne tient pas. Nous simulons maintenant l'absence de données conformément à l'hypothèse du modèle commun d'erreur de mesure décrit à la section 3. Nous utilisons le même modèle logistique avec les mêmes coefficients que pour la simulation précédente, excepté que y_i^* est remplacé par y_i^* à titre de covariable. Les estimations en simulation des biais

et des erreurs-types sont présentées au tableau 5. Nous observons un biais relatif significatif non négligeable d'environ 5 % pour l'approche d'imputation et un peu plus élevé pour l'approche de pondération par le score de propension. La direction positive du biais de θ_1 est conforme aux attentes fondées sur les arguments de Dickens et Manning (2004) et de Skinner et coll. (2002). Les méthodes basées sur l'hypothèse MAR auront tendance à surestimer le nombre de travailleurs faiblement rémunérés, si l'hypothèse d'erreur de mesure commune tient. Il en est ainsi parce que les employés pour lesquels les valeurs y_i^* sont observées ont tendance à être moins bien rémunérés que ceux pour lesquels les valeurs y_i^* manquent et qu'une méthode d'imputation fondée sur l'hypothèse MAR, même connaissant les autres variables, aurait tendance à imputer des valeurs plus faibles de rémunération horaire que cela ne serait le cas sous l'hypothèse d'erreur de mesure commune qui permet la dépendance par rapport à la rémunération horaire réelle. Bien que l'on puisse prévoir la direction de l'effet, la grandeur de celui-ci a une certaine importance en ce qui concerne la robustesse des méthodes fondées sur l'hypothèse MAR. Le biais relatif de 5 % de la méthode PPV10 ne semble toutefois pas rendre les estimations résultantes inutilisables.

Tableau 2

Estimations par simulation des erreurs-types des estimateurs de θ_1 et θ_2 pour diverses méthodes d'imputation, sous l'hypothèse de mécanisme MAR et de covariables correctes ($H = 1\ 000$)

Méthode d'imputation	$e\text{-}t(\theta_1)$	$e\text{-}t(\theta_2)$	$V^{Add}(\theta_1)$	$V^{Add}(\theta_2)$
PPV1	2,79*10 ⁻³	5,43*10 ⁻³	1	1
PPV1P ²	2,60*10 ⁻³	5,15*10 ⁻³	0,87	0,91
PPV2	2,68*10 ⁻³	5,05*10 ⁻³	0,91	0,86
PPV2(4)	2,73*10 ⁻³	4,88*10 ⁻³	0,94	0,80
PPV10	2,56*10 ⁻³	4,88*10 ⁻³	0,83	0,81
PPV10(20)	2,57*10 ⁻³	4,79*10 ⁻³	0,84	0,77
IHDR10	2,52*10 ⁻³	4,66*10 ⁻³	0,82	0,74
IHDRSR10	2,48*10 ⁻³	4,72*10 ⁻³	0,78	0,76
BBA10	2,63*10 ⁻³	4,87*10 ⁻³	0,88	0,80

² Nota : $H = 100$ itérations ont été utilisées à cause du temps de calcul.

Tableau 3

Estimations par simulation des biais des estimateurs de θ_1 et θ_2 pour l'imputation par le plus proche voisin (PPV10) et la pondération par le score de propension, sous l'hypothèse de mécanisme MAR et de covariables correctes et spécifiques incorrectement ($H = 1\ 000$)

Méthode	Covariables hypothétiques	Biais de θ_1	Biais rel. de θ_1	Biais de θ_2	Biais rel. de θ_2
PPV10	M1 (correct)	-0,18*10 ⁻⁴	-0,03 %	-5,8*10 ⁻⁴	-0,31 %
	M2	-1,31*10 ⁻⁴	-0,24 %	-4,74*10 ⁻⁴	-0,25 %
	M3	-1,66*10 ⁻⁴	-0,30 %	-10,6*10 ⁻⁴	-0,57 %
Pondération par le score de propension	M1 (correct)	0,15*10 ⁻⁴	0,03 %	-2,62*10 ⁻⁴	-0,14 %
	M2	-8,96*10 ⁻⁴	-1,64 %	70,2*10 ⁻⁴	3,80 %
	M3	-5,02*10 ⁻⁴	-0,92 %	67,8*10 ⁻⁴	3,66 %

M1 est le modèle correct.
M2 correspond à l'exclusion des termes d'interactions et des termes quadratiques du modèle correct.
M3 correspond à l'élimination des covariables supplémentaires par rapport au modèle correct.

car elle permet d'éviter le biais des méthodes basées sur des classes d'imputation et de réaliser des gains appréciables d'efficacité par rapport aux méthodes générant une ou deux imputations.

Nous allons maintenant comparer l'approche d'imputation PPV10 à la pondération par le score de propension (PSP). Nous considérons non seulement le cas où la spécification du modèle utilisé pour l'imputation ou pour la pondération correspond au modèle utilisé pour la simulation, comme au tableau 1, mais aussi certains cas de spécification incorrecte. Pour assurer que la comparaison de la pondération et de l'imputation soit équitable, nous utilisons les mêmes covariables lors de l'ajustement des deux modèles générant y_i et x_i . Nous considérons pour commencer les biais estimés présentés au tableau 3. Lorsque le modèle pour l'imputation (PPV10) ou pour la pondération par le score de propension est spécifié correctement, ni l'une ni l'autre méthode ne donne lieu à un biais significatif dans l'estimation de θ_1 ou θ_2 . Toutefois, nous observons un biais significatif dans les deux cas, si le modèle est mal spécifié en oubliant d'inclure les covariables utilisées dans la simulation. Néanmoins, le biais est sensiblement plus important pour l'approche de pondération. Par exemple, pour l'estimateur θ_1 , le biais est de 3 à 7 fois plus élevé sous la méthode PSP que sous la méthode PPV10, selon l'erreur de spécification. L'effet de l'erreur de spécification semble plus important pour l'estimateur θ_2 , en particulier sous la méthode PSP. Dans le cas de cet estimateur, nous observons un biais de 6 à 15 fois plus grand pour cette dernière que pour la méthode PPV10.

Tableau 1 Estimations par simulation des biais des estimateurs de θ_1 et θ_2 pour diverses méthodes d'imputation, sous l'hypothèse de mécanisme MAR et de covariables correctes ($H = 1\ 000$)

Méthode d'imputation	Biais de θ_1	Biais rel. de θ_1	Biais de θ_2	Biais rel. de θ_2
PPV1	$1,2*10^{-4}$	0,2 %	$0,9*10^{-4}$	0,0 %
PPV1P ¹	$(0,9*10^{-4})$	0,8 %	$(1,7*10^{-4})$	0,0 %
PPV2	$(2,6*10^{-4})$	0,1 %	$(5,1*10^{-4})$	0,0 %
PPV2(4)	$(0,8*10^{-4})$	0,2 %	$(1,5*10^{-4})$	-0,1 %
PPV10	$(0,9*10^{-4})$	0,0 %	$(1,3*10^{-4})$	-0,1 %
PPV10(20)	$(0,8*10^{-4})$	0,0 %	$(1,5*10^{-4})$	0,0 %
IHDAR10	$(0,8*10^{-4})$	0,5 %	$(1,5*10^{-4})$	1,4 %
IHDRS10	$(0,8*10^{-4})$	0,4 %	$(1,5*10^{-4})$	1,5 %
BBA10	$(0,8*10^{-4})$	0,8 %	$(1,5*10^{-4})$	1,6 %

¹ Nota : $H = 100$ itérations ont été utilisées à cause du temps de calcul.
Les erreurs-types des estimations du biais figurent entre parenthèses sous les estimations.

approches d'estimation de la variance sous imputation par le plus proche voisin, ainsi que Little et Rubin (2002) pour les approches d'imputation multiple.

7. Étude par simulation

Le but de l'étude est de générer des échantillons répétés indépendants $s^{(h)}, h = 1, \dots, H$, avec des valeurs $x_i, r_i, i \in s^{(h)}$ réalisées dans le contexte de l'application de l'EPA, examinée plus loin à la section 8, afin de calculer les estimations correspondantes $\hat{F}^{(h)}(y)$ pour diverses approches de traitement des valeurs manquantes de y et d'évaluer empiriquement les propriétés des estimateurs $\hat{F}(y)$. Afin d'utiliser des valeurs réalistes, nous avons tiré les échantillons $s^{(h)}$ de taille n avec remise (c'est-à-dire par la méthode du bootstrap) à partir d'un échantillon réel d'environ 16 000 employés pour le trimestre de mars à mai 2000 utilisé pour l'EPA (seuls les emplois principaux des employés de 18 ans et plus ont été pris en considération et le très petit nombre de cas pour lesquels des valeurs de y_i^* ou x_i manquaient ont été omis). Les valeurs de x_i pour chaque échantillon $s^{(h)}$ ont été tirées directement des valeurs dans l'échantillon de l'EPA. Les critères utilisés pour choisir les variables incluses dans x_i étaient qu'elles soient corrélées à la rémunération horaire, à l'erreur de mesure dans y_i^* ou à la réponse r_i (voir Skinner et coll. 2002). Ces variables comprennent, par exemple l'âge, le sexe, la position dans le ménage, les qualifications, la profession, la durée de l'emploi, le travail à temps plein/temps partiel, l'industrie et la région (plusieurs de ces variables ont été représentées par des variables muettes). Nous avons fixé $n = 15\,000$, de sorte que chaque $s^{(h)}$ soit de taille similaire à celle de l'échantillon original de l'EPA, et $H = 1\,000$. Les valeurs de y_i, y_i^* et r_i pour chaque échantillon $s^{(h)}$ ont été générées d'après des modèles, plutôt que directement d'après les données de l'EPA, pour les raisons qui suivent.

y_i^* : ces valeurs ont été générées d'après un modèle afin d'éviter les valeurs en double de (y_i^*, x_i) dans chaque échantillon $s^{(h)}$, ce qui, à notre avis, aurait pu donner lieu à une distribution irréaliste des distances entre les unités pour la méthode du plus proche voisin. Nous avons utilisé un modèle de

régression linéaire reliant $\ln(y_i^*)$ à x_i avec une erreur normale et avec 12 covariables, y compris un terme quadratique pour l'âge et un terme d'interaction, que nous avons ajusté aux données de l'EPA.

r_i : ces valeurs ont été générées d'après un modèle afin d'assurer que le mécanisme de production des données manquantes est MAR sachant les y_i^* et x_i pour tous les résultats présents, sauf ceux du tableau 5. Nous avons obtenu les estimations $\hat{\theta}_i^{(h)}$ de deux paramètres ($t = 1, 2$) pour chaque échantillon $s^{(h)}$.

θ_1 = proportion de travailleurs dont la rémunération est inférieure au salaire minimum national (= 3,00 £ de l'heure pour les personnes de 18 à 21 ans, 3,60 £ de l'heure pour les personnes de 22 ans et plus)

θ_2 = proportion de travailleurs dont le salaire minimum est compris entre le salaire minimum et 5 £ de l'heure.

Les valeurs réelles sont $\theta_1 = 0,056$ et $\theta_2 = 0,185$. Nous avons estimé le biais et l'erreur-type sous les formes

$$\text{biais}(\hat{\theta}_i) = \bar{\theta}_i - \theta_i, \text{ et } \hat{\sigma}_i = H^{-1} \sum_{h=1}^H (\hat{\theta}_i^{(h)} - \bar{\theta}_i)^2$$

Dans le cas des méthodes d'imputation fractionnaire, nous avons examiné plusieurs valeurs de M et avons choisi $M = 10$ ou 20 afin d'obtenir un accroissement de l'efficacité, tout en restant capable de définir raisonnablement une imputation par le plus proche voisin.

Nous commençons par comparer les résultats pour les diverses approches d'imputation. Le tableau 1 donne les estimations des biais des estimateurs de θ_1 et θ_2 pour diverses méthodes d'imputation, sous un mécanisme de données manquantes MAR. Nous ne dégageons aucune preuve d'un biais important pour aucune des méthodes par le plus proche voisin (PPV). Les ratios biais/erreur-type sont faibles et nous pouvons nous attendre à ce qu'ils soient encore plus petits pour les estimations pour des données tels que les régions ou les groupes d'âge. Nous concluons qu'il n'existe aucune preuve d'un biais important pour ces méthodes, à condition que l'hypothèse d'un mécanisme MAR tienne et que le modèle soit spécifié correctement.

Nous dégageons certaines preuves de l'existence d'un biais statistiquement significatif pour chacune des trois méthodes fondées sur les classes d'imputation (IHDSR10, IHDSR10, BBA10) peut-être dû à la largeur des classes, quoique le biais semble faible comparativement à l'erreur-type. Étant donné l'inconvénient supplémentaire que

Enfin, considérons l'effet des écarts par rapport à l'hypothèse de mécanisme MAR. Sous des conditions asymptotiques de faible erreur de mesure, où $y_i^* \rightarrow y_i$ et $V(u_i | z_i) \rightarrow 0$, de sorte que $y_i^* \rightarrow y_i$, l'approche d'imputation donne une inférence convergente au sujet de θ . même si l'hypothèse de mécanisme MAR ne tient pas. Par contre, ce n'est pas le cas pour l'approche de pondération par le score de propension. Cela donne à penser que l'approche d'imputation pourrait être plus robuste aux écarts par rapport à l'hypothèse MAR si l'erreur de mesure est relativement faible.

6. Estimation de la variance

Bien que l'estimation ponctuelle soit le sujet principal du présent article, nous allons maintenant considérer brièvement l'estimation de la variance par linéarisation. Dans le cas de la pondération par le score de propension, nous nous référons aux travaux de Kim (2004). Pour les méthodes d'imputation simple ou fractionnaire fondées sur l'imputation par le plus proche voisin décrites à la section 3, nous pouvons considérer une approche simplifiée basée sur l'hypothèse IID formulées à la section 2 et l'expression de la variance de $\hat{\theta}_{\text{IMP}}$ dans (13).

L'estimateur simple du premier terme σ^2/n :

$$n^{-1}\hat{\sigma}^2 = n^{-2} \sum_{i=1}^n w_i(n_i - \hat{\theta}_{\text{IMP}})^2 \quad (20)$$

est approximativement sans biais d'après le corollaire 1 de Chen et Shao (2000). Il s'ensuit qu'un estimateur approximativement sans biais de $\text{var}(\hat{\theta}_{\text{IMP}})$ est

$$V(\hat{\theta}_{\text{IMP}}) = n^{-1}\hat{\sigma}^2 + n^{-2} \sum_{i=1}^n (w_i^2 - w_i) V(u_i | z_i). \quad (21)$$

si nous pouvons construire un estimateur approximativement sans biais $V(u_i | z_i)$ de $V(u_i | z_i)$. Diverses approches d'estimation de $V(u_i | z_i)$ semblent possibles. À l'instar de Fay (1999), nous pourrions prendre en considération la variance d'échantillon de n_i valeurs pour les voisins répondants proches de i par rapport à z_i . Une autre approche serait d'adopter une méthode fondée sur un modèle selon laquelle un modèle est ajusté à $\psi(z_i) = E(n_i | z_i)$ pour $i \in s$ sachant $\psi(z_i)$ et de fixer $V(u_i | z_i) = \psi(z_i)[1 - \psi(z_i)]$. Nous avons envisagé des méthodes non paramétriques d'ajustement de $\psi(z_i)$, mais nous avons constaté qu'avec les données de l'EPA, elles mènent à des valeurs de $V(\hat{\theta}_{\text{IMP}})$ fort semblables à celles produites par un modèle de régression logistique pour $\psi(z_i)$.

Il pourrait être possible d'appliquer les idées de Chen et Shao (2001) ou de Kim et Fuller (2002) en vue d'étendre l'approche susmentionnée de façon à pouvoir tenir compte des poids de sondage et du plan de sondage complexe. Consulter Rancourt (1999) et Fay (1999) pour d'autres

$$\text{var}(\hat{\theta}_{\text{PS}}) \approx n^{-1}\hat{\sigma}^2 + n^{-2}E\left[\sum_{i=1}^n (w_i^2 - w_i) V(u_i | z_i) \right] + n^{-1}E\left\{\sum_{i=1}^n [w_i^2 - w_i] \psi(z_i) - [\psi(z_i)]^2\right\}. \quad (18)$$

Notons que, comparativement à (13), cette expression contient un troisième terme, qui ne converge pas nécessairement vers zéro quand $y_i^* \rightarrow y_i$ et que $V(u_i | z_i) \rightarrow 0$. Donc, la pondération par le score de propension ne devient pas entièrement efficace quand l'erreur de mesure disparaît. Nous pouvons aussi nous attendre à ce que les variances des coefficients de pondération w_i et w_{PSi} , et que, à condition que M soit suffisamment grand, nous pouvons nous attendre à ce que w_i soit moins variable que w_{PSi} , comme nous l'avons soutenu plus haut.

$$\text{var}(\hat{\theta}_{\text{PS}}) \approx n^{-2}E\left[\sum_{i=1}^n w_i^2 V(u_i | z_i) \right] + n^{-1}E\{\psi(z_i) - [\psi(z_i)]^2\}. \quad (17)$$

La discussion qui précède ne tient pas compte de l'effet éventuel de l'estimation de β ou de l'estimation d'un vecteur de paramètre α dont on peut supposer que dépend le score de propension $\Pr(y_i^* = x_i | y_i^*, x_i)$. Kim (2004) montre, en fait, que l'estimation de α par son estimateur du maximum de vraisemblance à réduit la variance de $\hat{\theta}_{\text{PS}}$ comme suit :

$$\text{var}(\hat{\theta}_{\text{PS}}) \approx \text{var}(\hat{\theta}_{\text{PS}}) - \text{cov}(\hat{\theta}_{\text{PS}}, \hat{\alpha}) \text{var}(\hat{\alpha})^{-1} \text{cov}(\hat{\alpha}, \hat{\theta}_{\text{PS}}). \quad (19)$$

où $\hat{\theta}_{\text{PS}}$ est l'estimateur $\hat{\theta}_{\text{PS}}$ dans lequel les scores de propension estimés sont remplacés par leur valeur réelle et où le premier membre de (16), (17) et (18) devrait maintenant être $\text{var}(\hat{\theta}_{\text{PS}})$. Nous concluons de ce fait et de la discussion qui précède qu'en général, $\hat{\theta}_{\text{IMP}}$ n'est pas forcément plus efficace que $\hat{\theta}_{\text{PS}}$ ou inversement, et nous appuyons sur l'étude en simulation présentée à la section 7 pour des preuves numériques. Cependant, la conclusion que $\hat{\theta}_{\text{IMP}}$ devient plus efficace à mesure que l'erreur de mesure disparaît et que $y_i^* \rightarrow y_i$ demeure valide, même en présence d'une erreur d'estimation dans α et dans β , puisque l'effet de l'erreur d'estimation dans β disparaît dans ce cas lorsque $z_i^* \rightarrow y_i^*$, tandis que le deuxième terme de (19), lorsqu'il est ajouté à l'expression (18), ne réduira généralement pas $\text{var}(\hat{\theta}_{\text{PS}})$ à σ^2/n dans ce cas.

proche voisin par rapport à z_i . Comme dans (9), l'estimateur imputé de θ peut être exprimé sous la forme

$$\hat{\theta}_{\text{IMP}} = \sum_{i \in s_1} w_i n_i / \sum_{i \in s_1} w_i \quad (10)$$

où $w_i = 1 + d_i / M$ (et $\sum_{i \in s_1} w_i = n$). Nous écrivons l'expression correspondante pour la pondération par le score de propension sous la forme θ_{PS} avec w_i remplacé par $w_{\text{PS},i}$. Représentons par $z_{\text{PS},i}$ la fonction scalaire de y_i^* , x_i dont dépend r_i et écrivons :

$$\Pr(r_i = 1 | y_i^*, x_i) = \pi(z_{\text{PS},i}). \quad (11)$$

Tout comme nous avons ignoré l'écart entre β et β_i , nous ignorons au départ l'erreur lors de l'estimation de $\pi(z_{\text{PS},i})$ et écrivons $w_{\text{PS},i} = \pi(z_{\text{PS},i})^{-1}$.

Nous pouvons nous attendre à ce que les approches d'imputation et de pondération par le score de propension produisent des estimateurs semblables si z_i et $z_{\text{PS},i}$ sont semblables, c'est-à-dire s'ils sont proches de fonctions déterministes l'un de l'autre, et que M est grand. Pour le montrer, considérons un exemple simple de l'approche d'imputation, où le donneur est tiré aléatoirement dans une classe d'imputation c de voisins proches par rapport à z_i , contenant m_c répondants et $n_c - m_c$ non-répondants, comme il est décrit à la section 3. Dans ce cas, w_i s'approchera de $1 + (n_c - m_c) / m_c = n_c / m_c$ lorsque $M \rightarrow \infty$ et l'agit de l'inverse du taux de réponse dans la classe (David, Little, Samuhan et Titter 1983). De façon plus générale, si nous suivons l'approche d'imputation fractionnaire par le plus proche voisin envisagée à la section 3, le poids $w_i = 1 + d_i / M$ peut être interprété comme une estimation locale (par rapport à z_i) non paramétrique de $\Pr(r_i = 1 | z_i)^{-1}$, malgré le fait que l'imputation est basée sur un modèle pour y_i sachant z_i plutôt que pour r_i sachant z_i . Donc, nous pouvons nous attendre à ce que la méthode d'imputation mène à des résultats d'estimation semblables à la méthode de pondération par le score de propension si z_i et $z_{\text{PS},i}$ sont des fonctions déterministes l'un de l'autre. Puisque $\Pr(r_i = 1 | z_i)$ peut être exprimé comme une moyenne de $\Pr(r_i = 1 | y^*, x)$ sur les valeurs de y^* et x pour lesquels $z = z_i$, nous pouvons interpréter w_i comme étant une version lissée de $w_{\text{PS},i}$ et pouvons nous attendre à ce que sa dispersion soit plus faible. Cela donne à penser qu'il pourrait être possible d'utiliser l'imputation pour améliorer l'efficacité des estimations fondées sur la pondération par le score de propension, comme en ont déjà discuté David et coll. (1983) et Rubin (1996, section 4.6). Pour étudier davantage cette possibilité, émettons l'hypothèse MAR, ainsi que les autres hypothèses formulées aux sections 3 et 4 sur lesquelles les approches sont fondées, de sorte que les approches d'imputation et de pondération mènent toutes

deux à une estimation approximativement sans biais de $F(y)$ et que nous puissions donc nous concentrer sur la comparaison des efficacités relatives. Il découle de l'équation (3.3) de Chen et Shao (2000) que la variance de $\hat{\theta}_{\text{IMP}}$ peut être approximée, pour une grande valeur de n , par

$$\text{var}(\hat{\theta}_{\text{IMP}}) \approx n^{-2} E \left[\sum_{i \in s_1} w_i^2 A(u_i | z_i) + n^{-1} A[\psi(z_i)] \right], \quad (12)$$

$$\text{var}(\hat{\theta}_{\text{IMP}}) \approx n^{-2} \sigma^2 + n^{-2} E \left[\sum_{i \in s_1} w_i^2 A(u_i | z_i) \right], \quad (13)$$

en utilisant l'identité

$$A[\psi(z_i)] = \sigma^2 - E[A(u_i | z_i)]. \quad (14)$$

où $\sigma^2 = A(u_i)$ et un corollaire du théorème 1 de Chen et Shao (2000) selon lequel

$$E \left[n^{-1} \sum_{i \in s_1} w_i^2 A(u_i | z_i) \right] = E[A(u_i | z_i)] + o_p(n^{-1/2}). \quad (15)$$

Notons que $w_i^2 - w_i = (d_i / M)(1 + d_i / M) \geq 0$. L'expression (13) peut être interprétée sous l'angle des « données manquantes » ainsi que sous celui de l'« erreur de

mesure ». Du point de vue des données manquantes, le premier terme de (13) est simplement la variance de $\hat{\theta}$ en l'absence de données manquantes et le deuxième terme représente l'inflation de cette variance due à l'erreur d'imputation. Du point de vue de l'erreur de mesure, nous pouvons considérer les propriétés limites sous les « conditions asymptotiques de faible erreur de mesure » (Chesher 1991), c'est-à-dire quand $y_i^* \rightarrow y_i$ et que $A(u_i | z_i)$ s'approche de zéro. Dans ce cas, le deuxième terme tend aussi vers zéro et $\hat{\theta}_{\text{IMP}}$ devient « entièrement efficace », c'est-à-dire que sa variance s'approche de σ^2 / n .

Considérons maintenant la pondération par le score de propension. Nous formulons l'hypothèse correspondante que y_i^* manque au hasard sachant $z_{\text{PS},i}$. En linéarisant le ratio de l'expression (9), avec $w_{\text{PS},i}$ à la place de w_i , en utilisant le fait que $E(\sum_{i \in s_1} w_{\text{PS},i}^2) = n$ et en ignorant au départ l'effet de l'estimation du score de propension, nous pouvons

$$\text{var}(\hat{\theta}^{\text{PS}}) \approx n^{-2} \text{var} \left[\sum_{i \in s_1} w_{\text{PS},i}^2 (u_i - \theta) \right] = n^{-1} E[w_{\text{PS},i}^2 (u_i - \theta)^2], \quad (16)$$

que nous pouvons aussi exprimer sous la forme

d'autres. Nous considérons plusieurs approches en vue de réduire cet effet d'inflation de la variance.

En premier lieu, nous pouvons restreindre le nombre de

fois que des répondants sont utilisés comme donneurs en

définissant des classes d'imputation au moyen d'intervalles

disjoints de valeurs de y_i et en tirant des donneurs pour un

receveur donné par échantillonnage aléatoire simple dans la

classe dans laquelle est comprise la valeur y_i du receveur.

Le lissage sera le plus important si nous tirons les donneurs

sans remise. Nous dénotons cette méthode hot deck IHDSR

ou IHDSR, selon que l'échantillonnage est fait avec ou sans

remise. Une deuxième approche consiste à sélectionner les

donneurs séquentiellement et à pénaliser la fonction de

distance $d(i)$ employée pour déterminer le plus proche

voisin comme suit

$$|y_i - y^{ad(i)}| = \min_{j: r_j = i} \{ |y_i - y_j| + (1 + \mu r_j) \}, \quad (8)$$

où $\mu \in \mathbb{R}^+$ est un facteur de pénalité, r_j est le nombre de

fois que le répondant j a déjà été utilisé comme donneur,

$r_j = 0$ et $r_j^{ad(i)} = 1$ (Kallion 1983). Une troisième approche

consiste à employer des valeurs imputées répétées $y_i^{(m)}$,

$m = 1, \dots, M$, pour chaque receveur $i \in s$, telles que $r_i =$

0. L'estimateur résultant de $F(y)$ est $M^{-1} \sum_{m=1}^M F^{(m)}(y)$, la

moyenne des estimateurs résultants $F^{(m)}(y)$. Nous dé-

nommons cette troisième approche imputation fractionnaire

(Kallion et Kish 1984; Fay 1996) plutôt qu'imputation

multiple (Rubin 1996), parce qu'il n'est pas nécessaire que

notre méthode d'imputation soit « appropriée », c'est-à-dire

qu'elle remplisse les conditions assurant que l'estimateur de

la variance par imputation multiple soit convergent. Nous ne

stipulons pas cette exigence ici, parce que notre objectif

premier est l'estimation ponctuelle. Lorsque nous utilisons

l'imputation fractionnaire, nous visons à sélectionner des

donneurs $d(i)$, $m = 1, \dots, M$ qui sont chacun un voisin

proche de i , de sorte que $F^{(m)}(y)$ demeure approximative-

ment sans biais pour $F(y)$. Nous examinons les va-

riantes suivantes de cette approche.

i) Les $M/2$ voisins les plus proches au-dessus et

au-dessous de la valeur y_i sont tirés, pour $M = 2$

ou 10, et dénotés PPV2 et PPV10, respectivement.

ii) $M/2$ donneurs sont sélectionnés par échantillon-

nage aléatoire simple avec remise parmi les M

répondants pour lesquels la valeur est supérieure à

y_i et parmi les M répondants pour lesquels elle est

inférieure, pour $M = 2$ ou 10, et nous les dénotons

PPV2(4) et PPV10(20), respectivement.

iii) $M = 10$ donneurs sont sélectionnés par échantillon-

nage aléatoire simple avec ou sans remise dans les

classes d'imputation auxquelles nous avons fait

allusion dans les méthodes IHDSR et IHDSR

5. Propriétés des approches d'imputation et de pondération

$$F(y) = \sum_{i \in s_i} w_i I(y_i < y) / \sum_{i \in s_i} w_i \quad (9)$$

exprime sous la forme pondérée :

L'estimateur $F(y)$ implique dans les diverses approches

d'imputation envisagées à la section précédente peut être

4. Estimation pondérée

des méthodes IHDSR et IHDSR.

Aux fins de comparaison, nous envisageons aussi la

méthode bootstrap bayésienne approximative d'imputation

multiple (Rubin et Schenker 1986), dénotée BBA10, définie

par rapport aux classes d'imputation mentionnées au sujet

des méthodes IHDSR et IHDSR.

décrites plus haut. Nous nommons ces méthodes

IHDAR10 et IHDSR10.

À la présente section, nous investiguons et comparons les

propriétés théoriques de l'approche d'imputation et de

l'approche de pondération par le score de propension pré-

sentées aux deux sections précédentes sous diverses hypo-

thèses simplificatrices. Nous fixons y et établissons $u_i =$

$I(y_i < y)$. Posant que $N \rightarrow \infty$, nous supposons que le

paramètre d'intérêt est $\theta = E(u_i)$. Nous considérons l'ap-

proche d'imputation pour commencer et supposons que y_i

proche d'imputation pour commencer et supposons que y_i

dépend de y_i^* et x_i uniquement par la voie de

$z_i = g^{-1}[h(y_i^*, x_i; \beta)]$ et que y_i manque au hasard, sachant

z_i . En ignorant la différence entre β et β_i , et en supposant

que s_i est grand, nous considérons l'imputation par le plus

Cependant, puisque y_i n'est observé que si $r_i = 1$, les données ne fournissent aucun renseignement direct au sujet de $f(y_i | r_i = 0)$ si nous n'émettons pas d'hypothèses supplémentaires. Nous considérons deux hypothèses possibles.

Hypothèse (MAR) : r_i et y_i sont conditionnellement indépendants sachant y_i^* et x_i .

Hypothèse (modèle commun d'erreur de mesure) : r_i et y_i^* sont conditionnellement indépendants sachant y_i et x_i .

La première hypothèse est celle faite classiquement lorsque l'on recourt à l'imputation ou à la pondération (Little et Rubin 2002) et celle que nous formulons ici. La deuxième hypothèse est celle voulant que le modèle d'erreur de mesure, défini comme étant la distribution conditionnelle de y_i^* sachant y_i et x_i , soit le même pour les répondants ($r_i = 1$) que pour les non-répondants ($r_i = 0$). Nous utiliserons la deuxième hypothèse pour l'étude en simulation à la section 7 afin d'évaluer la robustesse des méthodes fondées sur le mécanisme MAR. Cependant, sous la deuxième hypothèse, l'inférence est plus difficile et semble nécessiter des hypothèses de modélisation plus fortes au sujet de la distribution de y_i et x_i ; nous étudions ce problème dans le cadre d'autres travaux et ne poursuivons pas son examen ici. La vraisemblance de ces deux hypothèses pour l'application de l'EPA est examinée plus en détail dans Skinner et coll. (2002).

Sous l'hypothèse de mécanisme MAR, nous avons $f(y_i | y_i^*, x_i, r_i = 0) = f(y_i | y_i^*, x_i, r_i = 1)$ et une condition suffisante pour que $\bar{F}(Y)$ donne une estimation sans biais de $F(Y)$ est que

$$f(y_i^* | y_i^*, x_i, r_i = 0) = f(y_i^* | y_i^*, x_i, r_i = 1). \quad (4)$$

Par conséquent, nous envisageons une approche d'imputation où la distribution conditionnelle de y sachant y^* et x est « ajustée » aux données sur les répondants ($r_i = 1$), puis les valeurs imputées y_i^* sont « tirées » à partir de cette distribution ajustée, aux valeurs y_i^* et x_i observées pour les non-répondants. Supposons que la distribution conditionnelle $f(y_i | y_i^*, x_i, r_i = 1)$ puisse être représentée par un modèle de régression paramétrique tel que :

$$g(y_i^*) = h(y_i^*, x_i; \beta) + e_i, E(e_i | y_i^*, x_i) = 0 \quad (5)$$

où $g(\cdot)$ et $h(\cdot)$ sont des fonctions données et β est un vecteur de paramètres de régression. Un prédicteur ponctuel de y_i^* , étant donné un estimateur $\hat{\beta}$ de β basé sur les données des répondants, est

$$\hat{y}_i = g^{-1}[h(y_i^*, x_i; \hat{\beta})]. \quad (6)$$

Toutefois, l'utilisation de \hat{y}_i pour l'imputation peut entraîner une sous-estimation importante de $F(Y)$ pour les faibles valeurs de y_i , puisque une simple imputation par la

régression de ce genre devrait, en principe, réduire la variance de $F(Y)$ artificiellement (Little et Rubin 2002, page 64). Cet effet pourrait être évité en prenant $y_i^* = g^{-1}[h(y_i^*, x_i; \beta) + \varepsilon_i]$, où ε_i est un résidu empirique sélectionné aléatoirement (Little et Rubin 2002, page 65). Néanmoins, selon notre expérience, cette approche ne permet pas de générer des valeurs imputées qui reproduisent les « pics » des distributions de la rémunération horaire dans notre application et peut donner lieu à un biais autour de ces pics. Nous préférons par conséquent nous limiter aux méthodes d'imputation par donneur, où l'on fixe $y_i^* = y_i^{(a)}$ ($r_i = 0$) pour un répondant donneur $j = d(i)$ pour lequel $r_j = 1$. La valeur imputée d'après le donneur sera toujours une valeur authentique et respectera les pics de la distribution dans notre application. La méthode fondamentale d'imputation par donneur que nous considérons ici est l'imputation par appariement d'après la moyenne prévisionnelle (Little 1988), c'est-à-dire l'imputation par la méthode du plus proche voisin par rapport à y_i^* , définie par (6), c'est-

$$\text{imputer } y_i \text{ par } y_i^{d(i)} \quad \text{en satisfaisant } |\hat{y}_i - \hat{y}_i^{d(i)}| = \min_{j: r_j=1} |y_i - y_j| \quad (7)$$

$$\text{où } r_j = 0 \text{ et } r_j^{d(i)} = 1.$$

Le corollaire 2 du théorème 1 de Chen et Shao (2000) nous donne alors la justification théorique de l'absence d'approximative de biais dans l'estimateur résultant $\bar{F}(Y)$ de $F(Y)$, si les quatre contraintes suivantes sont vérifiées : i) y_i manque au hasard (MAR) sachant $z_i = g^{-1}[h(y_i^*, x_i; \beta)]$, où $\beta = \text{plim}(\hat{\beta})$, ii) l'espérance conditionnelle de y_i sachant z_i est monotone et continue en z_i , iii) les troisième moments de z_i et $E(y_i | z_i)$ sont finis et iv) la probabilité de réponse sachant z est bornée au-dessus de zéro. Ces contraintes semblent plausibles à condition que l'hypothèse MAR susmentionnée tienne, que la distribution de y_i dépende uniquement de y_i^* et x_i par la voie de z_i et que y_i^* soit une approximation raisonnablement bonne de y_i . En outre, le résultat de Chen et Shao (2000) doit être adapté au fait que le plus proche voisin est défini par rapport à \hat{y}_i , tandis que les contraintes susmentionnées sont énoncées par rapport à β . Cette adaptation semble vraisemblable puisque, pour un nombre suffisamment grand de répondants, les voisins proches par rapport à $\hat{y}_i = g^{-1}[h(y_i^*, x_i; \hat{\beta})]$ devraient également être des voisins proches par rapport à $z_i = g^{-1}[h(y_i^*, x_i; \beta)]$.

Des fondements théoriques sous-tendent la notion que l'imputation selon la méthode du plus proche voisin par rapport à \hat{y}_i donnera un estimateur approximativement sans biais de $F(Y)$, sous l'hypothèse MAR et certaines autres conditions plausibles. Il est également intéressant de considérer l'efficacité de $\bar{F}(Y)$. La variance de $\bar{F}(Y)$ pour l'imputation par le plus proche voisin pourrait être exagérée si l'on utilise certains donneurs plus fréquemment que

Dans l'application de l'EPA, les unités sont les employés, s est l'ensemble d'unités répondantes dans l'échantillon de l'EPA, y_i^* est la valeur de la variable dérivée de données pour faire une inférence au sujet de $F(y)$.

où $I(.)$ est la fonction de vérité ($I(E) = 1$ si E est vrai et $= 0$ autrement) et y peut prendre toute valeur spécifiée. Supposons qu'une enquête soit réalisée auprès d'un échantillon $s \subset U$ et que la variable soit mesurée sous la forme y_i^j pour les unités $i \in s$. La différence entre y_i^j et y_i représente l'erreur de mesure. Supposons que la valeur vraie y_i soit enregistrée pour un sous-ensemble d'unités échantillonnées et que nous écrivions $\eta_i = 1$ si y_i^j est enregistré et $\eta_i = 0$ autrement. Soit x_i un vecteur de variables auxiliaires également enregistrées durant l'enquête. Nos données comprennent les valeurs de y_i^j , x_i et η_i pour les unités $i \in s$ et les valeurs y_i^j pour les unités $i \in s$ quand $\eta_i = 1$. Le problème est de savoir comment utiliser ces

$$(I) \quad \langle \chi \rangle_I \sum_{I-N}^{\infty} = \langle \chi \rangle_F$$

Soit y_i la valeur (vraie) d'une variable d'intérêt associée à l'unité i dans une population finie U . La fonction de distribution de la variable dans U est :

2. Le problème d'estimation

La présentation de l'article est la suivante. À la section 2, nous discutons du problème de l'estimation. Aux sections 3 et 4, nous exposons les approches par imputation et par pondération, respectivement. À la section 5, nous étudions et comparons leurs propriétés du point de vue théorique, tandis qu'à la section 7, nous le faisons par étude en simulation. À la section 6, nous nous penchons brièvement sur l'estimation de la variance. À la section 8, nous discutons de l'application des méthodes à l'EPA. Enfin, à la section 9, nous présentons certaines conclusions.

mesure la remémoration horaire de façon nettement plus précise que la variable dérivée. Néanmoins, le problème de la variable directe est que les données manquent pour environ 43 % des cas. L'application est décrite dans les grandes lignes à la section 8 et exposée plus en détail dans Skinner et coll. (2002), qui proposent eux aussi de recourir à l'imputation pour régler le problème de l'erreur de mesure. Le présent article prolonge ces travaux en envisageant une plus grande gamme d'approches de traitement des données manquantes et en comparant leurs propriétés du point de vue théorique, ainsi que par simulation. L'approche d'imputation élaborée dans le présent article, qui étend celle considérée par Skinner et coll. (2002), est maintenant appliquée par l'Office for National Statistics du Royaume-Uni comme nouvelle méthode de production d'estimations de la

Une condition suffisante pour que $F^{\varepsilon}(y)$ soit un estimateur sans biais de $F^{\varepsilon}(y)$ est que la distribution conditionnelle de y_i' sachant $y_i = 0$, détermine $\|F^{\varepsilon}(y) - F(y)\|_1$ la même que la distribution conditionnelle de y_i' sachant $y_i = 1$.

$$(E) \quad (\lambda > 1) I \sum_{u=1}^l u = (\lambda)_{\sim}^I$$

sera un estimateur sans biais de $F(V)$, en ce sens que $E[F(V) - (F(V))] = 0$ pour tout V , où nous écrivons $s = \{1, \dots, n\}$ et où l'espérance est obtenue par rapport au modèle, sachant l'échantillon sélectionné s . Pour résoudre le problème dû au fait que y_i manque quand $n_i = 0$, nous supposons que y_i est remplacé dans (2) par une valeur imputée y_i^* quand $n_i = 0$ (et $i \in s$) et que $\hat{y}_i = y_i$ si $n_i = 1$ et $y_i^* = y_i^*$ autrement. L'estimateur résultant de $F(V)$ est

$$(2) \quad (\lambda > 1) I \sum_{u=1}^{\lambda-1} u = (\lambda)_{\frac{\lambda-1}{2}} H$$

Supposons au départ qu'il est possible d'observer y_i pour tout $i \in s$. Alors, sous les hypothèses énoncées à la section qui précède,

3. Approches d'imputation

$F(y)$ aux deux sections qui suivent.

Nous discutons de l'inférence dans un cadre fondé sur un modèle, dans lequel nous nous proposons que les valeurs de population $(y_i, y_i^*, x_i, x_i^*, i \in I, \dots)$ sont indépendantes et de même loi (IID) et que l'échantillonnage est ignorable, c'est-à-dire que la distribution de (y_i, y_i^*, x_i, x_i^*) est la même que $i \in s$ ou non. À la section 8, nous expliquerons comment la méthode élaborée sous ces hypothèses peut être adaptée aux conditions du plan d'échantillonnage de l'EPA et au recours à la pondération pour tenir compte de la non-réponse totale durant l'enquête.

données manquantes, à savoir :

- l'imputation de y_i pour les unités $i \in s$ où $r_i = 0$, en utilisant les valeurs y_i^* et x_i comme données auxiliaires;
- la pondération d'un estimateur fondé sur le sous-échantillon de répondants $s_1 = \{i \in s; r_i = 1\}$, en particulier, l'utilisation de la pondération par le score de propension (Little 1986).

Nous discuterons de ces approches de l'estimation de

directe pour l'employé i . Nous supposons que la valeur y_i^t de la variable rémunération horaire et y_i^t est la valeur de certaines valeurs d'intérêt principalement est l'absence de certaines valeurs y_i^t et nous envisageons deux approches pour traiter ces

Utilisation de méthodes de traitement des données manquantes pour corriger l'erreur de mesure dans une fonction de distribution

Gabriele B. Durrant et Chris Skinner¹

Résumé

Nous examinons le recours à l'imputation et à la pondération pour corriger l'erreur de mesure dans l'estimation d'une fonction de distribution. Le problème qui a motivé l'étude est celui de l'estimation de la distribution de la rémunération horaire au Royaume-Uni au moyen de données provenant de l'Enquête sur la population active. Les erreurs de mesure causent un biais et le but est d'utiliser des données auxiliaires, mesurées avec précision pour un sous-échantillon, en vue de le corriger. Nous envisageons divers estimateurs ponctuels, fondés sur différentes approches d'imputation et de pondération, dont l'imputation fractionnaire, l'imputation par la méthode du plus proche voisin, l'appariement d'après la moyenne prévisionnelle et la pondération par le score de propension à répondre. Nous comparons ensuite ces estimateurs ponctuels d'un point de vue théorique et par simulation. Nous recommandons d'adopter une approche d'imputation fractionnaire par appariement d'après la moyenne prévisionnelle. Elle donne les mêmes résultats que la pondération par le score de propension, mais a l'avantage d'être légèrement plus robuste et efficace.

Mots clés : Imputation par donneur; imputation fractionnaire; imputation hot deck; imputation multiple; imputation par le score de propension.

1. Introduction

L'erreur de mesure peut donner lieu à une estimation biaisée des fonctions de distribution (Fuller 1995). Dans le présent article, nous examinons diverses approches en vue de corriger ce biais quand, en plus des observations sur l'échantillon de la variable mesurée incorrectement, on dispose de valeurs de la variable mesurée correctement pour un sous-échantillon. Si ce dernier est sélectionné aléatoirement, la situation se résume à un cas du problème bien étudié de l'échantillonnage double (par exemple Tenenbein 1970). Dans ces conditions, nous pouvons produire des estimations sans biais à partir du sous-échantillon unique-ment, mais utiliser les données concernant la variable substitut corrélée pour l'ensemble de l'échantillon afin d'accroître l'efficacité. Consulter, par exemple, Luo, Stokes et Sager (1998). Dans le présent article, nous supposons que le sous-échantillon n'est pas sélectionné selon un plan randomisé connu, mais plutôt selon un mécanisme de pro-duction de données manquantes inconnu. Nous émettrons simplement l'hypothèse que les données exactes sur la variable manquent au hasard (MAR pour *missing at random*) (Little et Rubin 2002), sachant que les variables sont mesurées sur l'échantillon complet. Nous disposons de certaines méthodes d'inférence pour résoudre ce problème si nous sommes prêts à formuler des hypothèses fortes concernant les paramètres de la distribution réelle (par exemple Buonaccorsi 1990) ou du modèle d'erreur de mesure (par exemple Luo et coll. 1998). Toutefois, nous ne pousserons pas plus loin leur examen, car nous supposons

que nous sommes en présence d'une application pour laquelle des hypothèses de ce genre sont irréalistes. La nouvelle du présent article tient plutôt au fait que nous considérons l'inférence sous ces conditions d'erreur de mesure comme étant un problème de données manquantes et que nous étudions l'application de méthodes d'imputation et de pondération décrites dans la littérature sur le traitement des données manquantes. Nous nous concentrerons sur le choix des meilleures méthodes en vue d'améliorer l'estimation ponctuelle de la fonction de distribution, en ce qui a trait au biais, à l'efficacité et à la robustesse aux hypothèses de modélisation. Nous ne nous pencherons que brièvement sur l'estimation de la variance.

La présente étude a été motivée par un effort en vue d'estimer la distribution de la rémunération horaire au Royaume-Uni (R.-U.) à l'aide de données provenant de l'Enquête sur la population active (EPA). L'EPA offre deux moyens de mesurer la rémunération horaire. La méthode classique consiste à recueillir des renseignements sur les gains et le nombre d'heures travaillées, puis à calculer la rémunération horaire d'après cette information. Nous appelons la variable dérivée de cette façon *variable dérivée de rémunération horaire*. Une méthode plus récente de détermination de la rémunération horaire consiste à demander directement aux répondants de déclarer quelle est cette dernière. Nous appelons la mesure résultante de la rémunération horaire la *variable directe*. Skinner et coll. (2002) décrivent, avec preuves empiriques à l'appui, de nombreuses sources d'erreur de mesure dans la variable dérivée et concluent, d'après leur étude, que la variable directe

Thompson, S.K., et Collins, L.M. (2002). Adaptive sampling in research on risk-related behaviors. *Drug and Alcohol Dependence*, 68, S57-S67.

Thompson, S.K., et Frank, O. (2000). Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépistage de liens. *Techniques d'enquête*, 26, 99-112.

Bibliographie

- Markov pour l'analyse des données associées à des modèles compliqués est fréquente en statistique. Les approches décrites ici sont inhabituelles en ce sens que les méthodes par chaîne de Markov sont appliquées à des populations réelles pour obtenir effectivement des données, qui peuvent être facilement analysées manuellement. En fait, on pourrait aller une étape plus loin et construire un modèle de graphe stochastique bayésien complexe de la population en utilisant des méthodes de Monte Carlo par chaîne de Markov de la façon classique pour l'analyse des données, ainsi que pour leur collecte.
- Les plans d'échantillonnage à marche uniforme ou ciblée sont utiles pour obtenir des échantillons de nœuds acceptés présentant certaines propriétés désirables en ce qui concerne la population, qui fournissent des estimateurs très simples des quantités de population ou qui pourraient fournir un échantillon initial pour un autre plan d'échantillonnage. Il convient de souligner que les nœuds qui ont été observés, puis « rejetés » sous les conditions du plan contiennent de faire effectivement partie des données. Leur valeur peut encore être intégrée dans les estimations, au besoin, en appliquant la méthode de Rao-Blackwell, une fois que la chaîne a atteint approximativement l'équilibre, mais dans ces conditions, le calcul des estimations est complexe.
- Une autre option consiste à utiliser des méthodes fondées sur un modèle, comme les méthodes d'estimation bayésiennes. En plus de la modélisation appropriée de la population par graphe stochastique, ces méthodes requièrent une procédure de sélection initiale ignorable, condition qui n'est généralement pas satisfaite sous sélections initiales biaisées par les valeurs ou les degrés de nœud, ou bien la modélisation adéquate de la procédure de sélection non ignorable dans les équations de vraisemblance. Les plans d'échantillonnage à marche ciblée produisant une loi asymptotique non corrélée à la procédure de sélection non ignorable et, donc, approximativement non corrélée aux valeurs ou aux degrés de nœud en dehors de l'échantillon pourraient fournir les sélections initiales pour un échantillon auquel les méthodes d'inférence basées sur un modèle pourraient ensuite être appliquées.
- ## Remerciements
- La présente étude a été financée par le National Center for Health Statistics, la National Science Foundation (DMS-9626102 et DMS-0406229) et les National Institutes of Health (R01-DA09872). Je tiens à remercier John Poterat et Steve Muth de m'avoir prodigué des conseils et permis d'utiliser les données provenant de l'étude de Colorado Springs.
- Birnbaum, Z.W., et Sirken, M.G. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. *Vital and Health Statistics*, Série 2, No. 11. Washington: Government Printing Office.
- Birn, S., et Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International Wide Web Conference*, Elsevier, 107-117.
- Chow, M., et Thompson, S.K. (2003). Estimation avec plans bayésienne. *Techniques d'enquête*, 29, 221-230.
- Felix-Medina, M.H., et Thompson, S.K. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics*, 20, 19-38.
- Frank, O. (1977). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.
- Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.
- Frank, O., et Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.
- Haslings, W.K. (1970). Monte-Carlo sampling methods using Markov chains and their application. *Biometrika*, 57, 97-109.
- Heckathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.
- Heckathorn, D.D. (2002). Respondent driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.
- Henzinger, M.R., Heydon, A., Mitzenmacher, M. et Najork, M. (2000). On near-uniform URL sampling. *Proceedings of the Ninth International World Wide Web Conference*, Elsevier, 295-308.
- Klov Dahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. Dans *The Small World*, (Ed. M. Kochen) Norwood, NJ: Ablex Publishing, 176-210.
- Lovász, L. (1993). Random walks on graphs: A survey. Dans *Combinatorics, Paul Erdős is Eighty*, (Eds. D. Miklós, D. Sós et T. Szőni), János Bolyai Mathematical Society, Keszthely, Hungary, 2, 1-46.
- Poterat, J.J., Woodhouse, D.E., Rothenberg, R.B., Muth, S.Q., Darrow, W.W., Muth, J.B. et Reynolds, J.U. (1993). AIDS in Colorado Springs: Is there an epidemic? *AIDS*, 7, 1517-1521.
- Rothenberg, R.B., Woodhouse, D.E., Poterat, J.J., Muth, S.Q., Darrow, W.W., Muth, J.B. et Reynolds, J.U. (1995). Social networks in disease transmission: The Colorado Springs study. Dans *Social Networks*, (Eds. R.H. Needle, S.G. Genser et R.T. Trotter) Drug Abuse, and HIV Transmission, NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 3-19.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Saigamik, M.J., et Heckathorn, D.D. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34, 193-239.
- Spren, M. (1992). Rare populations, hidden populations, and link-tracing designs: what and why? *Bulletin de Méthodologie Sociologique*, 36, 34-58.
- Statistique Canada, N° 12-001-XPB au catalogue

Dans le cas de la population empirique tirée de l'étude sur la transmission hétérosexuelle du VIH/Sida, les taux d'acceptation obtenus pour les divers plans d'échantillonnage sont donnés au tableau 9. Pour le plan à marche uniforme, le taux d'acceptation est de 62 %. Pour la marche selon la valeur de nud, où la probabilité limite est deux fois plus élevée pour les personnes à haut risque que pour celles à faible risque, le taux d'acceptation est de 60 %. Pour la marche selon le degré de nud, dans laquelle la probabilité limite est proportionnelle au degré+1, il est de 85 %. Enfin, pour la marche selon le degré, avec une unité ajoutée uniquement au degré des nuds isolés, il est de 88 %.

9. Discussion

Les plans d'échantillonnage à marche uniforme et à marche ciblée ont pour but de permettre de déterminer les probabilités limites de sélection d'après les données, afin de pouvoir les utiliser dans l'estimation. En outre, les probabilités limites sont choisies de sorte que certains types de nuds ou de caractéristiques de graphe puissent être sélectionnés de manière préférentielle. La dépendance à l'égard de la sélection initiale, qui peut ne pas être contrôlée, diminue pas à pas.

Les estimateurs utilisés dans le présent article avec les plans d'échantillonnage à marche uniforme et à marche ciblée peuvent être considérés comme des estimateurs fondés sur le plan de sondage. Le plan exact fondé sur les probabilités de sélection pourrait ne pas être connu, si les probabilités de sélection initiales sont inconnues, mais on utilise les probabilités de sélection stationnaires dans les estimateurs. À mesure qu'augmente la longueur de la chaîne, ces probabilités deviennent plus exactes et l'espérance des estimations se rapproche de la quantité de graphe fondée sur le plan de sondage est que certaines de leurs propriétés, comme l'absence de biais ou la convergence par rapport au plan, ne dépendent pas d'hypothèses fondées sur un modèle qui pourraient être incorrectes. Les estimations fondées sur le plan de sondage ont la qualité intéressante supplémentaire d'être très simples et faciles à comprendre et à expliquer, et elles peuvent même produire des données qu'il est possible de présenter sans analyse ou interprétation comme étant représentatives de caractéristiques importantes de la population d'intérêt dans son ensemble.

Dans le cas des marches uniformes et ciblées, l'une des questions pratiques importantes est celle du taux d'acceptation, c'est-à-dire la probabilité moyenne qu'un nud sélectionné provisoirement soit accepté. Les nuds sélectionnés provisoirement qui sont rejetés ne contribuent pas aux estimations simples. Dans le cas d'une population telle qu'Internet, pour laquelle les sélections provisoires et les décisions d'acceptation/rejet peuvent être automatisées et exécutées rapidement, le taux d'acceptation n'est pas nécessairement critique. L'échantillonnage se poursuit simplement jusqu'à ce qu'un nombre approprié de nuds soit accepté. Par contre, lors des études de populations humaines cachées, les tailles d'échantillon sont généralement faibles. Les membres de la population sont difficiles à atteindre et les interviews peuvent prendre beaucoup de temps. Toutefois, dans certaines études, la décision d'accepter ou de rejeter une unité d'après le degré sortant d'une personne sélectionnée provisoirement peut être prise assez rapidement au moyen d'une brève interview de filtrage. Il est malgré tout souhaitable de disposer d'une méthode d'échantillonnage dont le taux d'acceptation est aussi élevé que possible.

Les marches aléatoires sont caractérisées par une probabilité d'acceptation égale à un, mais n'ont généralement pas de probabilités limites connues ou contrôlées. Si l'on se représente la marche aléatoire sous-jacente comme le cheminement naturel, non contrôlé, au sein d'une population, alors on pourrait s'attendre à ce qu'une marche contrôlée ayant une loi limite proche de la marche aléatoire naturelle de la population produise un taux d'acceptation plus élevé qu'une marche contrôlée dont la loi limite diffère considérablement de cette marche aléatoire naturelle. Autrement dit, une marche contrôlée dont la loi stationnaire s'écarte peu de la loi de la marche aléatoire sous-jacente devrait nécessiter moins de modifications sous-jacentes pour produire un taux d'acceptation plus élevé que les probabilités stationnaires d'une marche aléatoire ordinaire dans un graphe non orienté à une seule composante sont proportionnelles au degré des nuds. Lorsqu'il existe plus d'une composante connectée, l'ajout du saut aléatoire est nécessaire pour assurer que chaque nud puisse être atteint, pour produire une loi stationnaire unique ne dépendant pas de la loi initiale et pour faire en sorte que les probabilités limites soient influencées par le degré des nuds, mais qu'elles n'y soient pas strictement proportionnelles. Même avec l'introduction du saut aléatoire et des probabilités d'acceptation induites, les marches ciblées produisant des probabilités stationnaires proportionnelles de nud pour approcher davantage de la loi naturelle de la

Les tableaux 3 à 6 résument les espérances et les erreurs quadratiques moyennes des estimateurs calculées pour les diverses stratégies d'après les 1 000 simulations exécutées en prenant l'ensemble de données de Colorado Springs comme population.

Tableau 3

Moyennes et erreurs quadratiques moyennes des moyennes d'échantillon des unités distinctes et moyennes tirage par tirage pour les marches aléatoires et les marches uniformes. Le plan de sondage comporte 24 marches se poursuivant chacune jusqu'à ce que 5 nœuds distincts soient inclus

Plan :	Marche	Marche	Marche
	aléatoire	aléatoire	uniforme
Estimateur :	Moyenne	Moyenne	Moyenne
d'échantillon	du tirage	du tirage	du tirage
moyenne	0,3008000	0,2994872	0,2423000
e.g.m.	0,007617465	0,007608868	0,002016378

Tableau 4

Moyennes et erreurs quadratiques moyennes pour les moyennes pondérées (estimateur par le ratio généralisé), en utilisant les unités distinctes dans chaque marche ou les sélections tirage par tirage pour les marches en fonction de 24 marches se poursuivant chacune jusqu'à ce que 5 nœuds distincts soient inclus

Plan :	Marche selon	Marche selon	Marche selon
	la valeur	le degré	le degré
Estimateur :	Unités	Unités	Unités
d'échantillon	du tirage	du tirage	du tirage
moyenne	0,1805114	0,2144555	0,225257
e.g.m.	0,002546966	0,001195507	0,001807981

Tableau 5

Moyennes et erreurs quadratiques moyennes pour les moyennes d'échantillon des unités distinctes et les moyennes tirage par tirage pour les marches aléatoires et les marches uniformes. Le plan comprend une marche se poursuivant jusqu'à ce que 120 nœuds distincts soient inclus

Plan :	Marche	Marche	Marche
	aléatoire	aléatoire	uniforme
Estimateur :	Moyenne	Moyenne	Moyenne
d'échantillon	du tirage	du tirage	du tirage
moyenne	0,3274083	0,3325171	0,2379333
e.g.m.	0,012004961	0,014902382	0,002442825

Tableau 6

Moyennes et erreurs quadratiques moyennes pour les moyennes pondérées (estimateur par le ratio généralisé), en utilisant les unités distinctes dans chaque marche ou les sélections tirage par tirage pour les marches selon la valeur et selon le degré. Le plan comprend une marche se poursuivant jusqu'à ce que 120 nœuds distincts soient inclus

Plan :	Marche selon	Marche selon	Marche selon
	la valeur	le degré	le degré
Estimateur :	Unités	Unités	Unités
d'échantillon	du tirage	du tirage	du tirage
moyenne	0,1652275	0,2254267	0,2404622
e.g.m.	0,003952703	0,001578039	0,002115518

Les tableaux 7 et 8 donnent la variance et l'espérance et des variances d'échantillon intramarche pour les plans à marche uniforme.

Tableau 7

Estimateur :	Moyenne d'échantillon	Moyenne tirage par tirage
Variance de l'estimateur :	0,001665709	0,001947796
E (variance inter-marches)	0,001584203	0,001919005
E (variance intramarche moyenne)	0,001515521	0,001231983

Tableau 8

Estimateur :	Moyenne d'échantillon	Moyenne tirage par tirage
Variance de l'estimateur :	0,001571384	0,002445194
E (variance intramarche moyenne)	0,001510515	0,001429126

Tableau 9

Taux d'acceptation pour les marches uniforme et ciblé dans la population empirique

Plan :	Marche	Marche
	uniforme	selon la valeur
Taux d'acceptation	0,62	0,60
		degré+1
		degré
		0,85
		0,88

8. Taux d'acceptation

Les principaux avantages des plans d'échantillonnage à

chaîne de Markov contrôlé, comme les marches uniformes et ciblées, sont les suivants : 1) ils permettent de connaître les probabilités limites de sélection d'après les données, de sorte que celles-ci peuvent être utilisées dans l'estimation, 2) les probabilités limites sont choisies de sorte que certains types de nœuds ou de caractéristiques des graphes puissent être sélectionnés de manière préférentielle, 3) les estimations sont fondées sur le plan d'échantillonnage, de sorte que certaines de leurs propriétés essentielles ne dépendent pas des hypothèses, qui pourraient s'avérer incorrectes, au sujet du graphe de population proprement dit et 4) à mesure que la longueur de la chaîne augmente, l'espérance des estimations a tendance à évoluer vers la quantité correspondante des graphes, même lorsque la loi de sélection initiale diffère de la loi limite. En outre, les plans à marche uniforme produisent un échantillon qui, sans pondération ni analyse, est au pied de la lettre « représentatif » à certains égards de l'ensemble de la population.

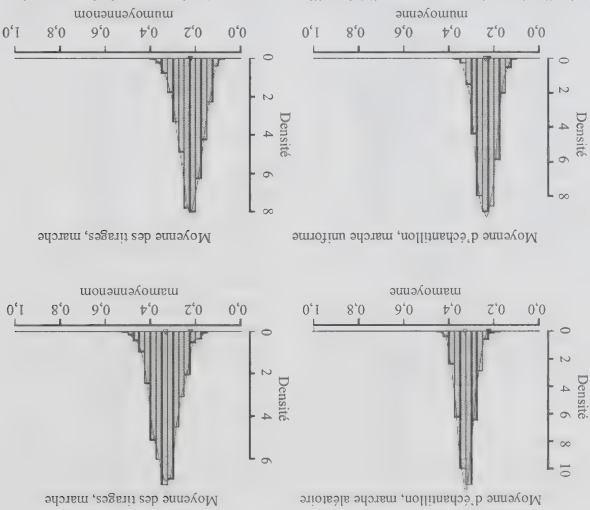


Figure 8.

Distributions des moyennes d'échantillon en tant qu'estimateurs de la proportion de personnes qui se sont prostituées dans la population empirique de l'étude de Colorado Springs, sous marches aléatoires et uniformes. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. À noter la surestimation dans le cas des marches aléatoires pour les marches aléatoires ordinaires. Les marches aléatoires sont représentées à la partie supérieure de la figure et les marches uniformes, à la partie inférieure. Le plan comprend une marche unique de 120 pas. Le nombre de réalisations de la simulation est égal à 1 000.

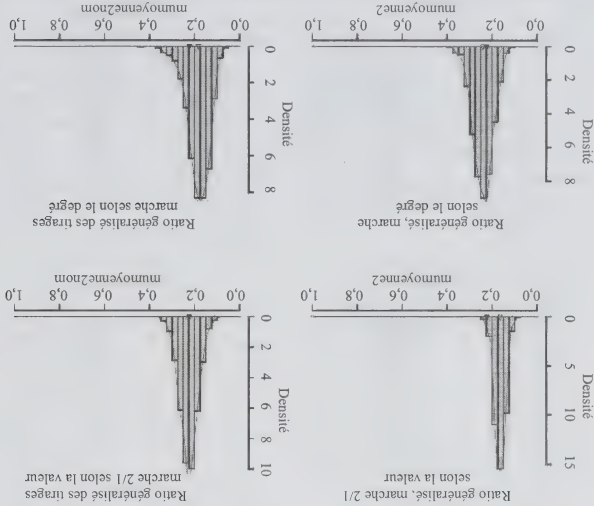


Figure 9.

Distributions des estimateurs par le ratio généralisé en tant qu'estimateurs de la proportion de personnes qui se sont prostituées dans la population empirique de l'étude de Colorado Springs, sous marches ciblées. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. À noter la surestimation dans le cas des moyennes d'échantillon pour les marches aléatoires ordinaires. Les marches aléatoires sont représentées à la partie supérieure de la figure et les marches uniformes, à la partie inférieure. Le plan comprend une marche unique de 120 pas. Le nombre de réalisations de la simulation est égal à 1 000.

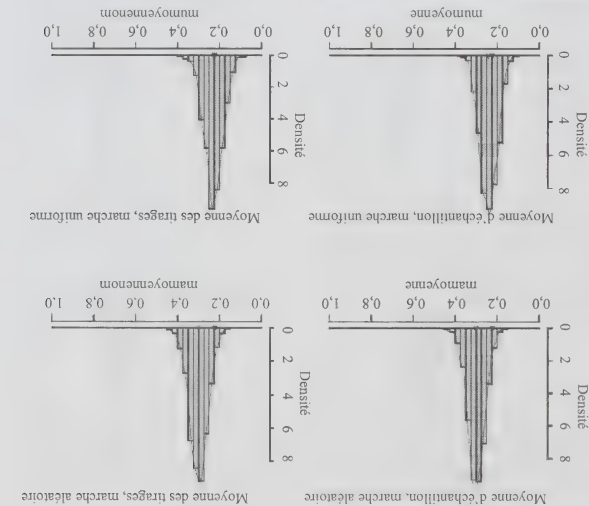


Figure 6.

Distributions des moyennes d'échantillon en tant qu'estimateurs de la proportion de personnes qui se sont prostituées dans la population empirique de l'étude de Colorado Springs, sous les marches aléatoires et uniformes. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. À noter la surestimation dans le cas des moyennes d'échantillon pour les marches aléatoires. Les marches aléatoires sont représentées à la partie inférieure. Le plan comporte 24 marches, chacune de 5 pas, et l'ensemble des 120 observations sont utilisées dans l'estimateur. Le nombre de réalisations de la simulation est égal à 1 000.

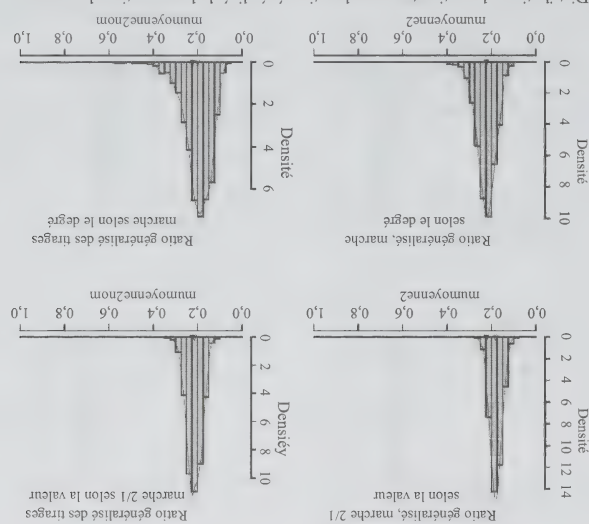


Figure 7.

Distributions des estimateurs par le ratio généralisé de la proportion de personnes qui se sont prostituées dans la population empirique de l'étude de Colorado Springs, sous marches ciblées. Le triangle plein représente la proportion réelle dans la population. Le triangle vide représente la moyenne de la distribution de l'estimateur. À noter la surestimation dans le cas des moyennes d'échantillon pour les marches aléatoires. Les marches aléatoires sont représentées à la partie supérieure de la figure et les marches uniformes, à la partie inférieure. Le plan comporte 24 marches, de 5 pas chacune, et l'ensemble des 120 observations sont utilisées dans l'estimateur. Le nombre de réalisations de la simulation est égal à 1 000.

selon un plan donné, les espérances à la vague j s'appliqueraient à la moyenne d'échantillon des k valeurs de y à la vague j provenant de chacune des marches.

Tableau 1

Marches aléatoires : espérance de y pour les vagues 0, 1, 2, 3, 4, 5, 6, 8, 16, 32 et à l'infini. La vague 0 correspond à la sélection initiale. Trois hypothèses de sélection initiale différentes sont appliquées : sélection initiale aléatoire ($\pi_0 = 1/N$ pour tous les nœuds), probabilité de sélection des nœuds de valeur $y = 1$ double de celle des nœuds de valeur $y = 0$ ($\pi_0 \propto y + 1$), et probabilité de sélection initiale proportionnelle au degré entrant du nœud plus 1 ($\pi_0 \propto a_j + 1$). La moyenne réelle des valeurs de nœud pour cette population est égale à 0,2235294

valeur	$\pi_0 = 1/N$	$\pi_0 \propto y + 1$	$\pi_0 \propto a_j + 1$
0	0,2235294	0,3653846	0,3349894
1	0,2998771	0,2752690	0,3560839
2	0,3005446	0,3587093	0,3507451
3	0,3273606	0,3082865	0,3570490
4	0,3177081	0,3594697	0,3500041
5	0,3320705	0,3179675	0,3528395
6	0,3231213	0,3542086	0,3469835
8	0,3356034	0,3440933	0,3440449
16	0,3291087	0,3372548	0,3363884
32	0,3302606	0,3313908	0,3315119
∞	0,3303787	0,3303787	0,3303787

Tableau 2

Marches uniformes : espérance de y pour les vagues 0, 1, 2, 3, 4, 5, 6, 8, 16, 32 et à l'infini, pour trois hypothèses de sélection initiale différentes

valeur	$\pi_0 = 1/N$	$\pi_0 \propto y + 1$	$\pi_0 \propto a_j + 1$
0	0,2235294	0,3653846	0,3349894
1	0,2235294	0,2590239	0,2903147
2	0,2235294	0,2741356	0,2877974
3	0,2235294	0,2447258	0,2761270
4	0,2235294	0,2511473	0,2707929
5	0,2235294	0,2372440	0,2646280
6	0,2235294	0,2420866	0,2600923
8	0,2235294	0,2371714	0,2522952
16	0,2235294	0,2285370	0,2352150
32	0,2235294	0,2243635	0,2256228
∞	0,2235294	0,2235294	0,2235294

le cas de la sélection initiale proportionnelle au degré entrant plus 1, quelques vagues de plus sont nécessaires pour que le biais devienne petit. Le rapprochement initial rapide de l'espérance vers la valeur limite donne à penser qu'il pourrait être souhaitable de considérer une période initiale « de rodage » qui ne sera pas utilisée dans l'estimation. Même un rodage très court d'une à trois vagues pourrait réduire sensiblement le biais des estimateurs fondés sur de courtes marches.

Les figures 6 à 9 illustrent les distributions d'échantillonage des moyennes d'échantillon et des estimateurs pondérés pour divers plans à marche aléatoire pour l'ensemble de données de Colorado Springs. Chaque histogramme est basé sur 1 000 simulations du plan d'échantillonage appliqué à la population empirique. Pour les plans illustrés aux figures 6 et 7, chaque échantillon comprend 24 marches, chacune de 5 pas, c'est-à-dire continuant jusqu'à ce que 5 nœuds distincts soient sélectionnés. La figure 5 représente les distributions des moyennes d'échantillon pour les marches aléatoires (rangée supérieure) et pour les marches uniformes (rangée inférieure). La distribution de la moyenne des 24 moyennes d'échantillon des 5 unités distinctes est donnée à gauche. À droite est donnée la moyenne des 24 moyennes tirage par tirage, qui intègre les sélections répétées.

La proportion réelle (0,2235) de nœuds ayant la valeur y pendant un nombre fixe de vagues.

La figure 7 illustre la distribution de l'estimateur par le ratio généralisé pour les marches ciblées dont les probabilités stationnaires sont reliées à la valeur des nœuds et à leur degré (degré du nœud plus 1). Aux fins de comparaison, chacune de ces marches a débuté dans sa propre loi stationnaire, ce qui donne en fait les distributions des estimateurs après le « rodage ». Ces estimateurs ne sont pas dépourvus de biais, parce que la taille effective de l'échantillon est fixe, ce qui affecte les probabilités réelles avec lesquelles les nœuds distincts sont sélectionnés en série et que le dénominateur de l'estimateur est aléatoire, puisqu'il est égal à la somme des poids de sondage.

Les figures 8 et 9 donnent les distributions des mêmes estimateurs et plans de sondage qu'aux figures 6 et 7, mais dans le cas où chaque échantillon consiste en une longue marche de 120 nœuds distincts.

nœud passablement plus élevée que la valeur moyenne dans la population. La marche « degré +1 » atteint une loi dont les probabilités de sélection sont proportionnelles à un plus la valeur de chaque nœud tandis que la marche selon le degré tend vers une loi dont les probabilités limites sont proportionnelles au degré de chaque nœud sauf que les nœuds isolés se voient attribuer une valeur de degré égale à un.

Les tableaux 1 et 2 montrent les valeurs calculées de l'espérance de γ pour la population de l'étude de Colorado Springs pour chaque type de marche, vague par vague, et avec diverses lois de départ pour la sélection des nœuds. Les résultats pour les marches aléatoires ordinaires sont présentés au tableau 1 et pour les marches uniformes, au tableau 2. Les espérances sont présentées pour les sélections initiales, les vagues 1, 2, 3, 4, 5, 6, 8, 16 et 32, et pour la limite à mesure que le nombre de vagues tend vers l'infini. Les trois lois initiales considérées pour la sélection du premier nœud d'une marche sont la sélection aléatoire, la sélection avec probabilité deux fois plus élevée pour les nœuds positifs que pour les nœuds à valeur nulle, et la sélection proportionnelle au degré entrant de chaque nœud plus 1. Notons que, dans le cas de k marches indépendantes

nuls ($\gamma = 0$), donne la courbe de l'espérance qui est, dans tous les cas, principalement au milieu au départ et manifeste la tendance la plus forte vers une périodicité initiale. La loi initiale fondée sur le degré, selon laquelle la probabilité initiale de sélection d'un nœud est proportionnelle à son degré (plus une unité, puisque le degré des nœuds isolés est nul), forme la courbe supérieure dans chacun des tracés.

Les six tracés de la figure 5 montrent les espérances des valeurs de nœud pour six différents types de marches. Pour une marche aléatoire qui suit seulement les liens, sans possibilité de sauts aléatoires, la loi à long terme est fonction du point de départ, lequel dépend de la loi initiale. Les trois lignes séparées du premier tracé reflètent la sensibilité à la loi initiale. Par ailleurs, la marche aléatoire avec sauts permet à n'importe quel nœud d'être atteint par n'importe quel autre de sorte que la loi limite est atteinte assez rapidement pour importer la loi initiale. Avec la marche uniforme, celle qui débute avec la loi uniforme demeure dans la loi uniforme, vague après vague tandis que les marches qui débutent avec les autres lois inégales décrites, tendent assez rapidement vers la loi uniforme. Chacune de marches qui dépendent de la valeur ou du degré atteint sa loi limite assez rapidement, avec une espérance de la valeur du

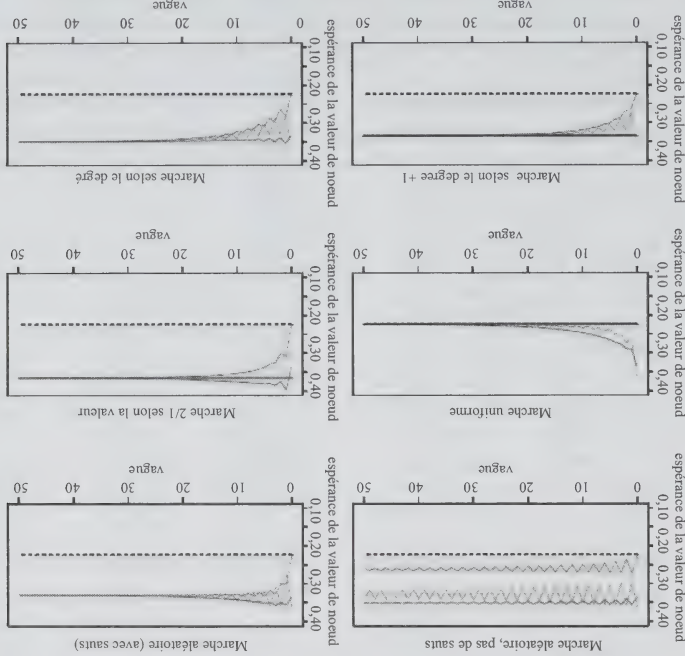


Figure 5.

Espérance de la valeur de nœud selon la vague pour divers plans de marche appliqués à la population empirique de Colorado Springs. Chaque tracé illustre un plan de marche. La courbe en trait interrompu représente la moyenne réelle. Les trois autres courbes représentent l'espérance pour les trois distributions initiales examinées. Dans chaque cas, la courbe inférieure débute par la loi uniforme, celle du milieu, par la loi de probabilité $2/1$ selon la valeur et la courbe supérieure, par la loi de probabilité selon le degré.

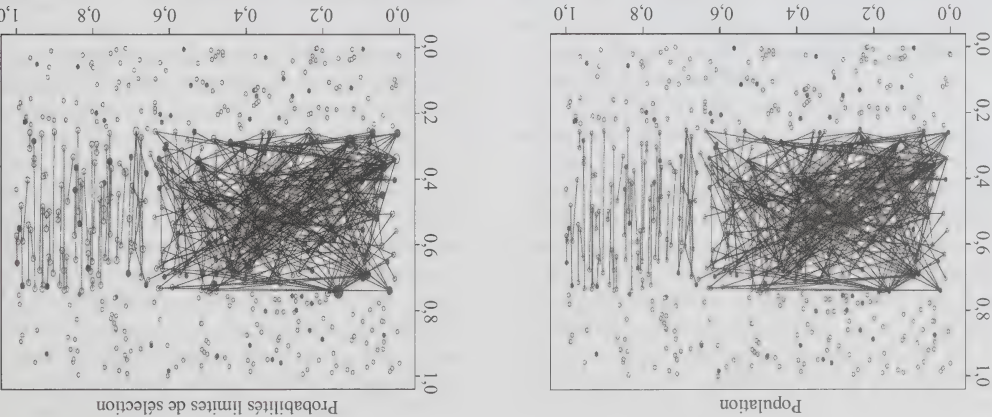


Figure 3.

Probabilités limites de sélection par marche aléatoire pour la population de Colorado Springs. Sou-
lignons que, dans la population réelle, un grand
nombre d'individus présentant le comportement à
risque le plus élevé ont aussi une forte probabilité
d'être sélectionnés dans le cas de la marche aléa-
toire ordinaire et auront donc tendance à être
sureprésentés dans l'échantillon.

Population à haut risque de l'étude de Colorado
Springs sur la transmission hétérosexuelle du
VIH/Sida (Poterat et coll. 1993; Rothenberg et coll.
1995, et communications personnelles). Les cercles
foncés représentent les individus présentant le risque
le plus élevé, ici ceux qui se sont prostitués. Les liens
entre les individus sont ceux de relations sexuelles et
d'injection de drogues.

Figure 2.

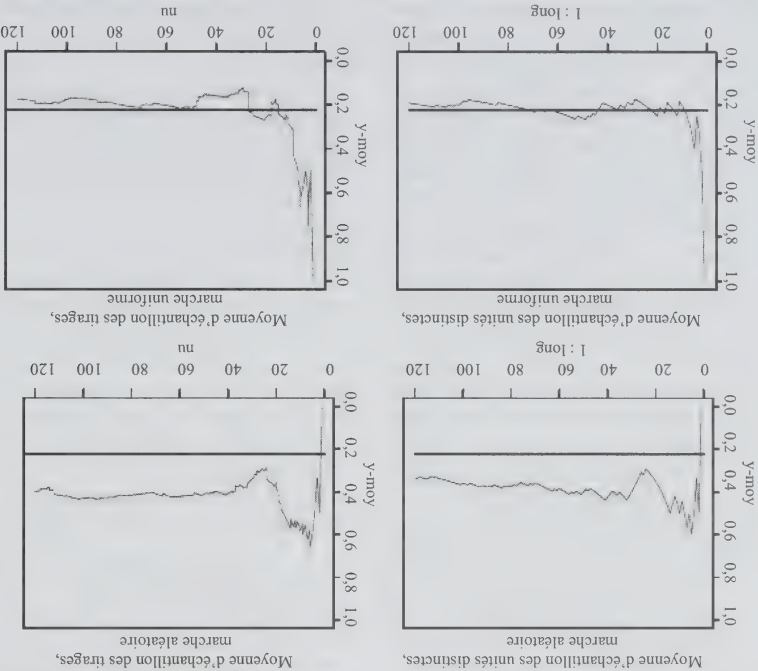


Figure 4.

Chemins d'échantillon des moyennes d'échantillon pour une marche aléatoire unique de 120 nœuds
de long. Les deux tracés supérieurs correspondent à une marche aléatoire ordinaire, et les deux tracés
inférieurs, à une marche uniforme. La moyenne d'échantillon des unités distinctes, jusqu'à la vague
donnée par l'axe des x, est représentée à gauche. La moyenne d'échantillon des tirages nominaux est
représentée à droite, de sorte que la valeur de nœud soit pondérée par le nombre de fois que le nœud
est sélectionné.

À la rangée inférieure de la figure 1 sont présentées une marche aléatoire et une marche uniforme sélectionnées à partir de la population, comme il est illustré. Chacune a pour point de départ le même nœud sélectionné au hasard, dénoté « 1 », et se poursuit jusqu'à ce que cinq nœuds distincts soient sélectionnés. Les flèches indiquent la direction dans laquelle sont suivis les liens et un saut vers un nouveau nœud sélectionné au hasard dans le graphique est indiqué par une ligne en pointillé. Notons que la marche aléatoire revient en arrière du troisième nœud sélectionné vers le deuxième avant de suivre un nouveau lien vers le quatrième nœud échantillonné. À partir du premier nœud échantillonné, la marche uniforme passe le nœud à probabilité plus élevée sélectionné par la marche aléatoire et accepte à sa place un des nœuds qui y sont liés. Ces marches peuvent l'une et l'autre, à tout moment, faire un saut aléatoire, quoique dans les exemples illustrés, seule la marche uniforme en fasse un, lors de la transition du troisième au quatrième nœud échantillonné.

Des données provenant d'une étude sur la transmission hétérosexuelle du VIH/Sida dans une population à risque élevée à Colorado Springs (Fortier et coll. 1993; Rothenberg et coll. 1995) sont présentées aux figures 2 et 3. Les 595 membres de la population étudiée qui ont été interviewés sont représentés par les nœuds du graphe, et les relations sexuelles déclarées entre ces personnes sont représentées par des liens (arcs) entre les nœuds. (Les liens d'ordre sexuel supplémentaires de n'importe laquelle de ces 595 personnes avec des personnes qui n'ont pas été interviewées subséquemment ne sont pas présents.) La population étudiée comprend les personnes à risque, c'est-à-dire les utilisateurs de drogues injectables, les travailleurs du sexe, leurs partenaires sexuels et d'usage de drogues, ainsi que d'autres personnes avec lesquelles ils ont des contacts sociaux étroits. La variable de nœud illustrée est celle de la prostitution, avec une couleur foncée pour une valeur positive ($y = 1$). Seuls les liens de nature sexuelle sont représentés, quoique bon nombre d'entre eux coïncident avec les liens relatifs à la consommation de drogues. La composante sexuellement connectée la plus importante du graphique contient 219 personnes. La composante connectée suivante, par ordre décroissant de taille, contient 12 personnes, et est suivie par plusieurs composantes de 4, 3 et 2 personnes. Les nœuds restants représentent des personnes n'ayant aucun contact sexuel déclaré au sein de la population interviewée.

Le profil observé de cette population, dont une composante connectée est beaucoup plus grande que les autres, a été décrit par divers chercheurs comme n'étant pas atypique des études portant sur des populations cachées à risque.

7.2 Population empirique

Nous utilisons la population susmentionnée uniquement à titre de population empirique à partir de laquelle nous sélectionnons des échantillons afin de comparer des plans d'échantillonnage et des estimateurs.

La figure 3 représente la même population avec la taille de nœud tirée proportionnellement à la probabilité de sélection limitée de la marche aléatoire.

Chaque tracé de la figure 4 montre la moyenne d'échantillon cumulative d'une marche unique poursuivie jusqu'à ce que 120 nœuds distincts soient sélectionnés. La portion réelle de nœuds positifs (valeur 1) dans la population empirique (0,2235) est représentée par la droite horizontale dans chaque tracé.

Les tracés de la rangée supérieure de la figure 4 représentent une marche aléatoire ordinaire dont le nœud de départ est sélectionné aléatoirement. Le tracé de gauche montre la moyenne d'échantillon cumulative des unités distinctes. Le tracé de droite montre les mêmes données, mais avec la moyenne d'échantillon tirage par tirage, qui comprend les sélections répétées d'un même nœud, de sorte que chaque valeur de nœud soit pondérée par le nombre de fois que le nœud a été sélectionné durant la marche aléatoire.

Dans la rangée inférieure de la figure 4, nous montrons les deux mêmes types de moyenne d'échantillon pour une marche uniforme poursuivie jusqu'à ce que 120 nœuds distincts soient sélectionnés. Notons que, pour la marche aléatoire ordinaire, les fluctuations de la moyenne d'échantillon ont lieu principalement au-dessus de la moyenne réelle, ce qui représente le biais positif résultant de la sélection préférentielle des personnes plus fortement connectées, à plus haut risque, dans la population. Dans le cas de la marche uniforme, la moyenne d'échantillon fluctue plus près de la valeur réelle, sa valeur étant parfois supérieure et parfois inférieure. Chacun de ces tracés donne aussi une idée de l'autocorrélation présente dans une chaîne de Markov unique.

Les tracés de la figure 5 représentent la valeur attendue des nœuds à mesure qu'une marche progresse vague par vague, pour divers types de marches et diverses lois initiales à partir desquelles est sélectionné le premier nœud, pour la population empirique de 595 nœuds. Donc, pour la k^{e} vague, les traces représentent $E(X_k)$, où X_k est la valeur du nœud sélectionné à la k^{e} vague. La ligne en trait interrompu montre la moyenne réelle pour la population de Colorado Springs (0,2235). Les trois autres courbes représentent trois lois initiales différentes. Dans tous les cas, la courbe qui part du nœud le plus bas est la loi initiale uniforme, puisque la moyenne pour le nœud initial sélectionné au hasard est égale à la moyenne de la population. La loi initiale fondée sur la valeur de nœud, selon laquelle les nœuds positifs ($y = 1$) ont une probabilité initiale de sélection double des nœuds

$$\hat{q} = \frac{\sum_{s_i} y_i / c_i}{\sum_{s_i} 1 / q_i}.$$

Notons que l'estimateur d'Horvitz-Thompson ne peut être utilisé, parce que la constante de proportionnalité des probabilités d'inclusion est inconnue, tandis que dans l'estimateur par le ratio généralisé, elle s'annule. De nouveau, les probabilités limites sur lesquelles est fondé l'estimateur sont vérifiées exactement pour le plan de sondage avec remise. Pour la variante sans remise, l'estimateur est examiné empiriquement dans les exemples.

7. Exemples

7.1 Graphe stochastique réalisé

La figure 1 illustre, pour commencer, une petite population simulée de 60 nœuds. Les nœuds dont la valeur est $y = 1$ sont de couleur foncée et ceux dont la valeur est

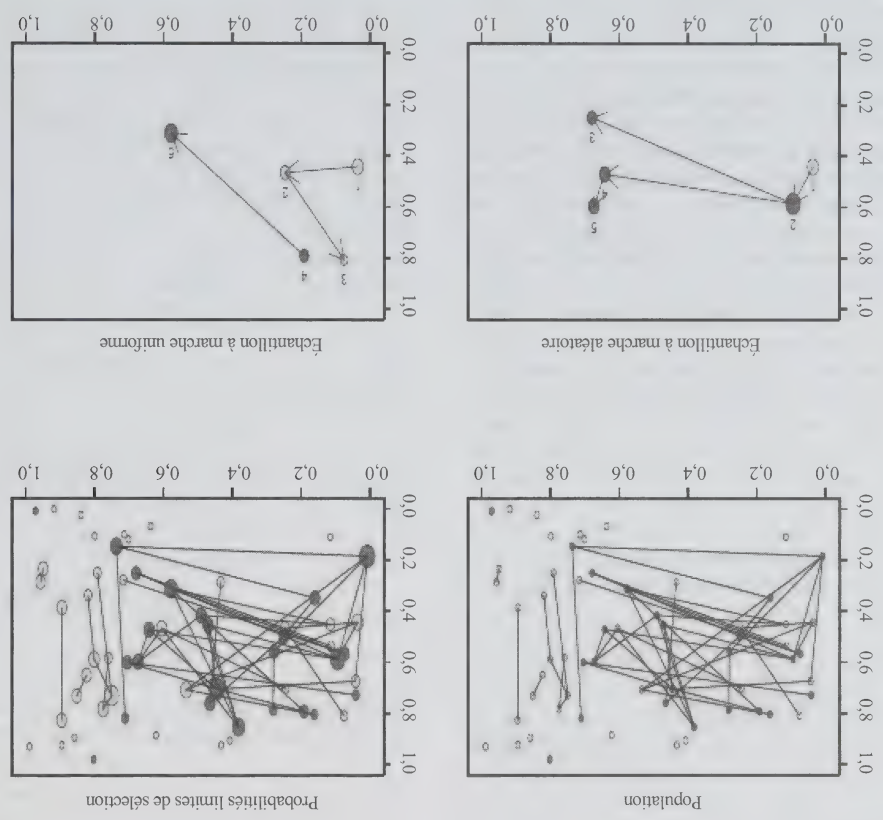


Figure 1. En haut à gauche : La population est la réalisation du modèle de graphe stochastique en blocs. En haut à droite : Probabilités limites de la marche aléatoire. En bas à gauche : Marche aléatoire : Marche aléatoire de cinq pas. En bas à droite : Marche uniforme de cinq pas. Des échelles d'axe arbitraires sont fournies comme aides visuelles pour distinguer les nœuds d'échantillon des nœuds de population.

5. Plan de sondage à marche sans remise

Les résultats relatifs aux lois limites des sections précédentes s'appliquent exactement au plan de sondage à marche aléatoire avec remise, de sorte que la sélection des nœuds peut se poursuivre indéfiniment au sein de la population finie. Certains estimateurs utilisés dans les exemples qui suivront sont toutefois fondés sur la séquence d'unités distinctes sélectionnées par ce processus. Dans le cas de la séquence d'unités distinctes, qui, en fait, fournit un échantillon à marche aléatoire sans remise, on ne peut ajouter des unités que jusqu'à ce que le nombre de nœuds distincts soit le même dans l'échantillon que dans la population finie, point auquel la moyenne d'échantillon et la moyenne de population coïncident.

Une autre procédure en vue de sélectionner un échantillon par marche aléatoire sans remise consiste à restreindre directement la sélection de l'unité suivante, à n'importe quel pas, à l'ensemble d'unités qui n'ont pas encore été sélectionnées, comme dans la « marche aléatoire auto-évitante » (Lovász 1993). Si l'on utilise une procédure sélection-reflet comme dans les marches ciblées, la sélection suivante est faite d'après l'ensemble d'unités qui n'ont fait l'objet d'aucune sélection provisoire, que l'unité ait été ou non acceptée.

6. Estimateurs fondés sur les valeurs des nœuds acceptés

Sous une marche aléatoire uniforme avec remise, la moyenne d'échantillon tirage par tirage de la série de valeurs acceptées est asymptotiquement sans biais par rapport à la moyenne de population, parce que les probabilités de sélection limites sont toutes égales. La moyenne de l'échantillon tirage par tirage est la moyenne nominale englobant les valeurs répétées, de sorte que la valeur d'un nœud est pondérée par le nombre de fois que le nœud est sélectionné. Si l'on utilise un plan de sondage sans remise, ce même estimateur n'est pas précisément asymptotiquement sans biais, parce que les probabilités limites ne sont pas exactement égales. L'estimateur de variance type fondé sur une variance d'échantillon à marche aléatoire n'est pas sans biais, à cause de l'interdépendance des marches aléatoires. Les estimateurs de la variance sont estimés empiriquement dans les exemples.

Dans le cas d'une marche ciblée dans laquelle la probabilité limite π_j du nœud j est proportionnelle à c_j , un estimateur asymptotiquement convergent, fondé sur les probabilités limites, est fourni par l'estimateur par le ratio généralisé

$$\alpha_{ij} = \min \left\{ \pi_j^{(Y_j)} q_{ij}^{(Y_j)}, \pi_j^{(X_j)} q_{ij}^{(X_j)} \right\}, 1$$

$\pi_j^{(Y_j)} / \pi_j^{(X_j)}$ est spécifié et

Un autre exemple de marche ciblée pourrait être celui d'une distribution cible obtenue en sélectionnant les nœuds

proportionnellement à leur degré, c'est-à-dire au nombre de liens qui en partent. Puisque le degré d'un nœud isolé est nul, une possibilité que nous nommerons marche ciblée selon le « degré + 1 », consiste simplement à ajouter une unité à chaque degré, de sorte que $\pi_j \propto a_j + 1$ soit la probabilité de sélection cible.

Un choix légèrement différent, appelé simplement marche ciblée selon le degré, consiste à n'ajouter une unité qu'au degré des nœuds isolés, de sorte que $\pi_j \propto \max(a_j, 1)$. Pour une marche ciblée selon le degré de ce type, la probabilité d'acceptation d'une transition entre deux nœuds connectés mutuellement est

$$\alpha_{ij} = \min \left\{ \frac{a_j(1 - p) / N + 1}{a_j(1 - p) / N + 1}, \frac{a_i(1 - p) / N + 1}{a_i(1 - p) / N + 1} \right\}$$

Pour une transition entre un nœud isolé et un nœud dont le degré est positif, la probabilité est

$$\alpha_{ij} = \min(a_j, (1 - p) / N)$$

La probabilité de transition entre deux nœuds ayant chacun un degré positif est

$$\alpha_{ij} = \min \left\{ \frac{a_i}{a_j}, 1 \right\}$$

Dans ce cas,

$$\alpha_{ij} = \min \left\{ \frac{a_j q_{ij}}{a_i q_{ji}}, 1 \right\}$$

Puisque les nœuds isolés, n'ayant aucun lien avec d'autres nœuds, sont de degré nul, afin de leur donner une probabilité de sélection positive, on peut attribuer arbitrairement à leur degré la valeur « 1 » dans le calcul de la marche ciblée selon le degré ou ajouter la valeur 1 au degré de chaque nœud.

d'échantillon (voir Henzinger et coll. 2000, pour une approche de ce problème).
 Dans la suite de l'article, les expressions « marche aléatoire » ou « marche aléatoire ordinaire » feront référence à la marche aléatoire avec sauts, sauf indication contraire explicite.

3. Marche uniforme

À la présente section, nous proposons une modification du plan d'échantillonnage à marche aléatoire qui mène à des probabilités stationnaires uniformes $\pi = (\pi_1, \dots, \pi_r)$.

Commençons par considérer le cas du graphe de population constitué d'une seule composante connectée.

Soit \mathcal{Q} la matrice de transition pour la marche aléatoire simple avec les probabilités de transition q_{ij} données par (1). Supposons qu'au pas k , l'état du processus est i . Une sélection provisoire est faite en utilisant les probabilités de transition de la i^{e} ligne de \mathcal{Q} . Supposons que la sélection provisoire soit le nœud j . Si le degré sortant a_j du nœud j est inférieur au degré sortant a_i du nœud i , alors la sélection pour la vague suivante est le nœud j , c'est-à-dire $W^{k+1} = j$. Si, par contre, le degré sortant du nœud j est supérieur au degré sortant du nœud i , alors un nombre aléatoire uniforme Z est sélectionné dans l'intervalle unitaire. Si $Z < a_i/a_j$, alors $W^{k+1} = j$. Sinon, $W^{k+1} = i$.

En utilisant la méthode de Hastings-Metropolis (Hastings 1970), nous construisons la matrice de transition pour la marche modifiée dans le graphe connecté au moyen des éléments

$$P_{ij} = q_{ij} \alpha_{ij} \quad \text{pour } i \neq j$$

et

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

où

$$\alpha_{ij} = \min \left\{ \frac{a_j}{a_i}, 1 \right\}.$$

Dans le cas d'un graphe de population contenant des composantes distinctes ou des nœuds isolés, la marche aléatoire avec sauts, dont la matrice de transition \mathcal{Q} est donnée par (2), peut être modifiée pour obtenir

$$P_{ij} = q_{ij} \alpha_{ij} \quad \text{pour } i \neq j$$

et

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

où

où

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij}$$

et

$$P_{ij} = q_{ij} \alpha_{ij} \quad \text{pour } i \neq j$$

sont

La même approche peut être suivie pour construire une marche ayant n'importe quelle probabilité stationnaire spécifiée, comme la sélection des nœuds dont la valeur y est élevée avec de plus grandes probabilités ou la sélection de nœuds de telle façon que les probabilités soient strictement proportionnelles à leur degré, même si le graphe contient des composantes distinctes connectées. Soit $\pi_i(y)$ la probabilité de sélection stationnaire souhaitée pour le i^{e} nœud sous forme d'une fonction de sa valeur de y . Par exemple, lors d'une étude d'une population humaine cachée exposée au risque de transmission du VIH/Sida, supposons que l'on souhaite échantillonner les utilisateurs de drogues injectables ($y_i = 1$) avec une probabilité double de celle appliquée pour les membres de cette population qui ne prennent pas ce genre de drogues ($y_i = 0$). Les probabilités de transition pertinentes pour la marche à valeur ciblée, en utilisant de nouveau la méthode de Hastings-Metropolis, sont

4. Marche ciblée

nécessite pas d'estimation.

La quantité α_{ij} , dans le cas des plans à marche uniforme, dépend des probabilités de transition connues de la marche aléatoire de base, si bien que sa mise en œuvre ne

La marche uniforme est appliquée, si l'état courant est i , en sélectionnant un prochain état candidat, disons j , d'après les probabilités de transition figurant sur la i^{e} ligne de \mathcal{Q} . Un nombre aléatoire uniforme standard Z est sélectionné et, si $Z < \alpha_{ij}$, l'état suivant est j , tandis qu'autrement, la marche reste à l'état i pendant un pas supplémentaire.

Pour une transition d'une unité isolée à une unité dans une composante plus grande qu'un nœud, la probabilité d'acceptation est $\alpha_{ij} = 1 - a$. Pour les autres probabilités d'acceptation, $\alpha_{ij} = 1$. Notons aussi que dans un graphe orienté, la probabilité d'acceptation serait nulle pour le parcours d'un lien asymétrique.

Donc, pour deux nœuds mutuellement connectés i et j , la probabilité d'acceptation d'une transition de i à j est

$$\alpha_{ij} = \min \left\{ \frac{(1-d)/N + d/a_i}{(1-d)/N + d/a_j}, 1 \right\}.$$

$$\alpha_{ij} = \min \left\{ \frac{q_{ji}}{q_{ij}}, 1 \right\}.$$

2. Marche aléatoire

La population d'intérêt est un graphe, donné par un

ensemble de N nœuds portant les étiquettes $U = \{1, 2, \dots, N\}$ et ayant les valeurs $y = \{y_1, \dots, y_N\}$, et une matrice \mathbf{A} de dimensions $N \times N$ indiquant les relations ou les liens entre les nœuds. Un élément a_{ij} de \mathbf{A} a la valeur 1 s'il existe un lien allant de i au nœud j et la valeur 0, autrement. Nous supposons que les éléments diagonaux a_{ii} sont nuls. Pour le nœud i , la somme de ligne $a_{i\cdot}$ est le « degré sortant » ou nombre de nœuds vers lesquels i possède un lien (successeurs) et la somme de colonne $a_{\cdot i}$ est le « degré entrant » ou nombre de nœuds qui ont un lien vers i (prédécesseurs). Dans le cas d'un graphe non orienté, la matrice \mathbf{A} est symétrique et le degré entrant de tout nœud est égal à son degré sortant.

Soit W_k l'unité ou le nœud du graphique qui est sélectionné lors de la k^{e} vague. Si i est le nœud sélectionné à la k^{e} vague, alors à la vague $k + 1$, l'un des nœuds reliés en partant de i est sélectionné au hasard. Donc, $\{W_0, W_1, W_2, \dots\}$ est une chaîne de Markov avec

$$(1) \quad P(W_{k+1} = j | W_k = i) = a_{ij} / a_{i\cdot}.$$

Soit \mathbf{Q} la matrice de transition de la chaîne avec les éléments $q_{ij} = P(W_{k+1} = j | W_k = i)$. La chaîne est une marche aléatoire en ce sens qu'à chaque pas, l'un des états voisins de l'état courant est sélectionné au hasard.

Si le graphe est constitué d'une seule composante connectée, c'est-à-dire si chaque nœud du graphe peut être atteint à partir de chaque autre nœud selon un certain chemin, alors la chaîne est irréductible et ses probabilités stationnaires (π_1, \dots, π_N) satisfont $\pi_j = \sum_i \pi_i q_{ij}$ pour $j = 1, \dots, N$. En fait, dans le cas du chaînonlancement à marche aléatoire simple d'un graphe non orienté connecté, on peut montrer que les probabilités stationnaires (Salganik et Heckathorn 2004) sont

$$\pi_j \propto a_{j\cdot}.$$

Autrement dit, dans un graphe non orienté ne comportant qu'une seule composante connectée, la fréquence de sélection de long terme de tout nœud est proportionnelle à son degré entrant, qui est égal au degré sortant, puisque le graphe n'est pas orienté.

Supposons que l'on veuille estimer une caractéristique du graphe de population, telle que la moyenne de population des valeurs de nœud $\bar{y} = \sum_{i=1}^N y_i / N$ en utilisant des données provenant d'un chaînonlancement par marche aléatoire. La moyenne d'échantillon $\bar{y} = \sum_{i=1}^n y_i / n$ est généralement pas sans biais, parce que la valeur y_i d'un nœud peut être reliée au degré de celui-ci et, donc, à sa probabilité d'être sélectionné. Cependant, on peut obtenir une estimation approximativement sans biais en pondérant chaque valeur y de l'échantillon par l'inverse de son degré

entrant, en supposant que cette information puisse être extraite des données (Salganik et Heckathorn 2004).

2.1 Marche aléatoire avec sauts aléatoires

Dans un graphe avec composantes distinctes ou avec nœuds non connectés, la marche aléatoire simple que nous venons de décrire n'a pas la propriété que chaque nœud peut, en dernière analyse, être atteint à partir de chaque autre nœud. Sans cette propriété, la loi limite de la marche aléatoire est sensible à la loi initiale, puisque la probabilité limite de sélection d'un nœud dépend de la probabilité initiale de démarrer dans la composante qui contient ce nœud. Une modification du plan d'échantillonnage qui permet de surmonter ce problème consiste à autoriser un saut avec faible probabilité vers un nœud choisi au hasard dans l'ensemble du graphe. À chaque pas, cette marche aléatoire suit un lien sélectionné au hasard avec la probabilité d et, avec la probabilité $1 - d$, saute à un autre nœud du graphe au hasard ou avec une probabilité spécifiée. Dans la littérature traitant de la recherche au sujet d'Internet, d est appelé « facteur d'amortissement », puisqu'une valeur de d inférieure à 1 amoortit l'effet du degré sortant d'un nœud donné (Brin et Page 1998).

Les probabilités de transition pour la marche aléatoire avec sauts sont

$$(2) \quad q_{ij} = \begin{cases} (1 - d) / N + d a_{ij} / a_{i\cdot}, & \text{si } a_{i\cdot} > 0 \\ 1 / N, & \text{si } a_{i\cdot} = 0. \end{cases}$$

Dans le cas de la faible probabilité $1 - d$ d'un saut aléatoire à n'importe quel pas, la marche aléatoire markovienne peut, en principe, atteindre tout nœud du graphe à partir de tout nœud, de sorte que la chaîne est irréductible. En outre, les sauts aléatoires, qui comprennent la possibilité d'aller du nœud i au nœud j , assurent que la chaîne soit aperiodique de sorte que les probabilités stationnaires concordent avec les probabilités limites. Si $d < 1$, la probabilité stationnaire du nœud i n'est pas une fonction simple de son propre degré entrant et dépend aussi des probabilités stationnaires des nœuds qui s'y relient.

De façon plus générale, les sauts peuvent être faits avec n'importe quelle probabilité spécifiée $\mathbf{p} = (p_1, \dots, p_N)$ et la probabilité d'un saut peut dépendre de l'état courant, de sorte que les probabilités de transition sont

$$q_{ij} = \begin{cases} (1 - p_j) d_j + p_j a_{ij} / a_{i\cdot}, & \text{si } a_{i\cdot} > 0 \\ 1 / N, & \text{si } a_{i\cdot} = 0. \end{cases}$$

Des estimations des caractéristiques du graphe de population approximativement sans biais par rapport au plan peuvent être obtenues en pondérant les valeurs d'échantillon par des facteurs inversement proportionnels aux probabilités limites de sélection de la chaîne de Markov, mais avec le problème supplémentaire que ces probabilités limites sont inconnues et doivent être estimées d'après les données

raisonnable (Rubin 1976) ou que le plan de sondage est de forme connue de sorte qu'il peut être inclus dans les équations de la vraisemblance et de l'inférence bayésienne, ces méthodes conviennent à une très grande gamme de procédures d'échantillonnage par dépiége de liens, y compris la plupart des variantes des méthodes d'échantillonnage en boule de neige et en réseau. Toutefois, en pratique, il se peut que l'échantillon initial soit sélectionné d'une façon loin d'être ignorable, avec probabilités de sélection dépendant de la valeur de neud, du degré de neud et d'autres facteurs. L'omniprésence du problème de la sélection de l'échantillon initial dans les études par dépiége de liens a été soulignée par Spreen (1992), entre autres.

L'approche poursuivie dans le présent article ne repose pas sur l'hypothèse d'un contrôle total sur toutes les possibilités de plan de sondage, mais vise plutôt à tirer parti de la façon dont les échantillons ont naturellement tendance à être sélectionnés dans les populations en réseau par les ethnographes ou d'autres spécialistes des sciences sociales, les membres de la population proprement dits ou les moteurs de recherche Web automatisés. Partant de ces processus naturels de sélection, nous introduisons des modifications itératives afin d'obtenir des procédures d'échantillonnage qui, pas à pas, s'approchent des probabilités de sélection souhaitées.

Quoique la structure sous-jacente des plans de sondage décrits dans l'article dépende de chaînes de Markov, les estimateurs et les paramètres présentant le plus d'intérêt pour les chercheurs pourraient en fait ne pas être markoviens. Par exemple, alors que la séquence de sélection d'unités d'échantillonnage peut ne dépendre, à chaque pas, que de l'unité sélectionnée le plus récemment, la séquence selon laquelle des unités distinctes sont ajoutées à l'échantillon dépend de toutes les unités sélectionnées jusqu'à ce moment-là. Par conséquent, nous étudions par simulation les propriétés de plusieurs estimateurs conjugués à divers plans de sondage, en sélectionnant répétitivement des échantillons à partir de réalisations d'un graphe stochastique et à partir d'une population empirique provenant d'une étude sur des personnes courant un grand risque de transmission du VIH/Sida.

À la section 2, nous décrivons les plans à marche aléatoire. Aux sections 3 et 4, nous présentons les plans à marche uniforme et ciblée, respectivement. À la section 5, nous donnons un exemple illustratif en prenant pour population une réalisation d'un modèle de graphe stochastique et un exemple empirique en utilisant des données provenant d'une étude portant sur une population présentant un risque élevé de transmission du VIH/Sida.

Le plan de sondage à marche aléatoire peut être conçu-
alisé comme une chaîne de Markov (Hecckathorn 1997, 2004). Dans le présent article, nous décrivons certains modifications apportées à ces plans de sondage à chaîne de Markov, dans le but d'obtenir des probabilités stationnaires de valeur égale ou spécifiée afin d'obtenir des estimations simples des caractéristiques du graphe de la population d'intérêt.

Les approches de l'inférence à partir d'échantillons provenant d'un graphe comprennent les méthodes fondées sur le plan de sondage, les méthodes fondées sur un modèle et les méthodes mixtes fondées sur une combinaison des deux. Dans l'approche fondée sur le plan de sondage, toutes les valeurs des variables de neud et de lien du graphe sont considérées comme étant fixes ou données, et l'inférence est basée sur les probabilités induites par le plan de sondage intervenant dans la sélection de l'échantillon. Dans l'approche fondée sur un modèle, la population proprement dite est considérée comme une réalisation d'un modèle de graphe stochastique, qui fournit la loi de probabilité conjointe de toutes les variables de neud et de lien. Les approches fondées sur le plan de sondage décrites antérieurement comprennent les méthodes d'échantillonnage en réseau ou basé sur la multiplicité (Birmbaum et Stikens 1965). L'échantillonnage en grappes adaptatif appliqué à un graphe (Thompson et Collins 2002), ainsi que quelques-unes des méthodes décrites dans la littérature sur l'échantillonnage en boule de neige (Frank 1977, 1978; Frank et Snijders 1994). Une méthode combinant les approches fondées sur le plan de sondage et sur un modèle est utilisée dans Felix-Medina et Thompson (2004) pour étudier une population cachée dans laquelle un échantillonnage par dépiége de liens est réalisé à partir d'un échantillon d'enquête probabiliste tiré d'une base de sondage couvrant uniquement une partie de la population.

L'avantage des méthodes fondées sur le plan de sondage, dans le cas de populations humaines cachées qui ont leur propre réseau social et sont difficiles à modéliser de façon réaliste, est que certaines propriétés des inférence, comme l'absence de biais et la convergence des estimateurs, ne dépendent pas d'hypothèses de modélisation. Par contre, elles dépendent de la mise en œuvre du plan de sondage comme il a été prévu; or, l'application exacte d'un plan de sondage particulier peut constituer un très grand défi dans les études de populations humaines cachées. C'est ce qui a motivé l'élaboration d'une gamme de méthodes fondées sur un modèle pour l'inférence à partir d'échantillons de graphe, y compris les techniques du maximum de vraisemblance et les techniques bayésiennes (Thompson et Frank 2000; Chow et Thompson 2003). Fondées sur l'hypothèse que l'échantillon de départ est « ignorable » au sens de la

Plans de sondage à marche aléatoire ciblée

Steven K. Thompson¹

Résumé

Les populations humaines cachées, Internet et d'autres structures en réseau conceptualisées mathématiquement sous forme de graphes sont intrinsèquement difficiles à échantillonner par les moyens conventionnels et les plans d'étude les plus efficaces comportent habituellement des procédures de sélection de l'échantillon par suivi adaptatif des liens reliant un nœud à un autre. Les données d'échantillon obtenues dans le cadre de telles études ne sont généralement pas représentatives au pied de la lettre de la population d'intérêt dans son ensemble. Cependant, un certain nombre de méthodes fondées sur le plan de sondage no sur un modèle sont maintenant disponibles pour faire des inférences efficaces à partir d'échantillons de ce type. Les méthodes fondées sur le plan de sondage ont l'avantage de ne pas s'appuyer sur un modèle de population hypothétique, mais dépendent, en ce qui concerne leur validité, de la mise en œuvre du plan de sondage dans des conditions contrôlées et connues, ce qui est parfois difficile, voire impossible, en pratique. Les méthodes fondées sur un modèle offrent plus de souplesse quant au plan de sondage, mais requièrent que la population soit modélisée au moyen de modèles de graphes stochastiques et que le plan de sondage soit ignorable ou de forme connue, afin qu'il puisse être inclus dans les équations de vraisemblance ou d'inférence bayésienne. Aussi bien pour les méthodes basées sur le plan de sondage que celles fondées sur un modèle, le point faible est souvent le manque de contrôle concernant l'obtention de l'échantillon initial, à partir duquel débute le dépistage des liens. Les plans de sondage décrits dans le présent article offrent une troisième méthode, dans laquelle les probabilités de sélection de l'échantillon deviennent pas-à-pas moins dépendantes de la sélection de l'échantillon initial. Un modèle de « marche aléatoire » markovienne idéalisée au moyen d'un graphe, les tendances d'un plan d'échantillonnage naturel d'une séquence de sélections par dépistage de liens à suivre. Le présent article présente des plans de sondage à marche uniforme ou ciblée dans lesquels la marche aléatoire est ajustée à chaque pas afin de produire un plan de sondage ayant les probabilités stationnaires souhaitées. On obtient ainsi un échantillon qui, à d'importants égards, est représentatif au pied de la lettre de la population d'intérêt dans son ensemble, ou qui ne nécessite que de simples facteurs de pondération pour qu'il en soit ainsi.

Mots clés : Échantillonnage adaptatif, échantillonnage déterminé selon les répondants (Respondent-driven sampling), échantillonnage d'une population cachée; échantillonnage en réseau; échantillonnage par graphes; marche aléatoire; méthode de Monte Carlo par chaîne de Markov; plans d'échantillonnage par dépistage de liens.

1. Introduction

Les populations comportant des liens ou une structure en réseau sont conceptualisées sous forme de graphes dans lesquels les nœuds (ou sommets) représentent les unités de la population et les arêtes ou les arcs, les relations ou liens entre ces unités. L'un des grands problèmes des études par établissement de graphes est qu'il est difficile, voire impossible, pour de nombreuses populations d'intérêt, d'obtenir des échantillons au moyen des plans de sondage conventionnels et que les échantillons sélectionnés peuvent être, tels qu'ils sont obtenus, fortement non représentatifs de la population d'intérêt dans son ensemble. En pratique, les seules méthodes d'échantillonnage applicables consistent souvent à suivre les liens à partir des nœuds sélectionnés, afin d'y ajouter des nœuds et des liens supplémentaires. Par exemple, lors de l'étude de populations humaines cachées, telles que les utilisateurs de drogues injectables, les travailleurs du sexe et d'autres populations courant le risque de contracter ou de transmettre le VIH/Sida ou l'hépatite C, les liens sociaux sont suivis en partant des répondants identifiés au départ, afin d'accroître l'échantillon de participants à

l'étude. De même, dans les études des caractéristiques d'Internet, la procédure habituelle consiste à obtenir un échantillon de sites Web en suivant les liens allant des sites initiaux vers d'autres sites. Klov Dahl (1989) a utilisé l'expression « marche aléatoire » pour décrire une procédure conçue afin d'obtenir un échantillon à partir d'une population cachée en demandant à un répondant d'identifier plusieurs contacts, dont un est sélectionné au hasard pour être le répondant suivant, et en répétant le scénario pendant un certain nombre de pas. Heckathorn (1997) a décrit des méthodes d'échantillonnage déterminé selon les répondants « répondent-driven sampling » en appliquant des procédures de ce genre. En pratique, la raison qui motive l'utilisation de plans de sondage de ce type est de pénétrer plus en profondeur dans la population cachée afin d'obtenir des répondants plus « représentatifs » de la population que ne le sont peut-être les personnes plus visibles sélectionnées initialement. Dans les études d'Internet, l'idée parallèle est que l'« internaute aléatoire », qui choisit une page Web au hasard, clique ensuite au hasard sur l'un des liens figurant sur cette page, passant ainsi à une autre page, et ainsi de suite (Brin et Page 1998).

En ses propres mots
Eric Rancourt
Statistique Canada

M.P. était un homme doté d'une impressionnante personnalité. Beaucoup de ses employés et collègues n'ont pas eu la chance de travailler étroitement avec lui, mais pour ceux dont ce fut le cas, il s'est révélé comme une personne très humaine et très polyvalente. Vous trouverez ci-dessous quelques-unes de ses citations que d'autres et moi-même avons rassemblées. C'était habituellement des phrases encouragantes qui nous incitaient à toujours repartir de son bureau du bon pied!

- Pas besoin d'une réunion, ma porte est toujours ouverte pour discuter de n'importe quoi.
- Il est bon d'avoir un projet de prédilection.
- Nous ne concevons pas les enquêtes pour calculer la variance.
- Je suis certain que cela peut se faire.
- Vous me dites que les deux tiers de vos suggestions ne se retrouvent pas dans l'enquête! Ne vous plaignez pas; si seulement 10 % de vos idées sont mises en application, vous aurez une carrière exceptionnelle.
- Il y a un panneau sur l'autoroute qui indique 100 km/h, cela ne veut pas dire qu'il faut rouler à 100 km/h.
- Ne vous inquiétez pas, il y a encore du temps.
- Après tout le travail que nous mettons à concevoir des enquêtes, ce que nous nous rappelons et ap-précions le plus ne sont pas les méthodes ni les résultats; ce sont les personnes avec lesquelles nous avons travaillé.

(Traductions libres).

gestionnaire non autoritaire qui savait encourager son personnel dans son travail. Même si M.P. pratiquait une gestion ouverte et souple, il savait quand faire acte d'autorité, comme beaucoup d'entre nous qui ont travaillé avec lui en ont fait la rude expérience, bien qu'en de rares occasions.

M.P. était un penseur stratégique qui aimait discuter en profondeur de statistique et de gestion. Cela donnait quelque-fois lieu à de longues réunions où nous devions tous donner notre point de vue. Au moment où nous pensions que le problème était réglé, M.P. intervenait et la discussion repartait! Evidemment, l'avantage de son approche était qu'à la fin de la réunion, nous comprenions tous les tenants et les aboutissants du sujet et que nous parvenions toujours à un consensus.

Tout au long de sa carrière, M.P. s'est toujours fortement intéressé au perfectionnement des chercheurs et à la recherche à Statistique Canada. Pour lui, un programme de recherche actif était essentiel au succès continu de Statistique Canada. C'est pourquoi il a travaillé à améliorer la visibilité profes-sionnelle des chercheurs et, d'une façon plus générale, des méthodologistes d'enquête, au sein de la Société statistique du Canada et dans d'autres organismes.

M.P., le superviseur et le mentor : après quelques années sous sa direction, j'ai eu la chance de l'avoir comme supérieur direct. Il est impossible de distinguer le superviseur et le mentor chez M.P. Il portait un réel intérêt à la carrière de ses employés immédiats, leur prodiguant des conseils et les orientant vers des choix judicieux ou, beaucoup plus im-portant, les orientant loin des mauvais choix. L'approche qu'il utilisait souvent est intéressante : au lieu d'être direct, il amenait souvent l'employé, d'une façon presque socratique, à comprendre que ce n'était pas une si bonne idée. Une autre technique était celle du « regard » – quiconque l'a bien connu savait d'un simple coup d'œil si M.P. considérait une idée comme particulièrement mauvaise.

M.P. m'a beaucoup appris au sujet des enquêtes mais plus important encore, j'ai appris de lui ce qui fait un bon gestion-naire, un bon motivateur et un bon mentor. Je conçois mainte-nant que son plus grand rôle était peut-être celui d'*enseignant*. Ceux d'entre nous qui ont travaillé étroitement avec lui au fil des ans continueront à profiter de son exemple pour le reste de leur carrière, et je m'attends à ce que nous légions à la prochaine génération ce que nous avons appris de lui après l'avoir intégré à notre propre expérience.

Singh, M.P. (1988). *Encyclopedia of Statistical Sciences*, (Eds. D.L. Banks, Read, B. Campbell et S. Kotz), New York : John Wiley & Sons, Inc. Vol. 9, 109-110.

Gestionnaire et mentor

Jack Gambino

Statistique Canada

D'autres ont écrit sur les contributions importantes et variées de M.P. Singh à la profession statistique et à Statistique Canada. J'ai eu la bonne fortune de travailler étroitement avec M.P. pendant 17 ans et de connaître plusieurs aspects de lui que seuls ceux qui le côtoyaient régulièrement ont vu et apprécié. J'ai vu M.P. dans son rôle de rédacteur en chef à *Techniques d'enquête*, dans ses activités quotidiennes qui menaient à la publication de chaque numéro de la revue, dans son rôle de gestionnaire et dans son rôle de superviseur et de mentor.

Dans les années 80, lorsque j'ai joint Statistique Canada, il était impossible de ne pas rencontrer M.P. Singh. Pendant les premières années, il était pour moi la personne qui posait les questions clés à tous les séminaires de méthodologie auxquels j'assistais. Beaucoup plus tard, lorsque nous siégeons ensemble à quelques comités, j'étais toujours fasciné lorsque, pendant les réunions, il posait de bonnes questions sur des sujets qui étaient clairement en dehors du domaine de la méthodologie. Invariablement, ses questions aidaient à clarifier les problèmes, non seulement pour les méthodologistes mais aussi pour tous les participants. Cela m'a fait comprendre que je ne devais pas présupposer que j'étais la seule personne à ne pas saisir complètement le sujet de discussion.

M.P., le rédacteur en chef : j'ai commencé à connaître personnellement M.P. lorsque j'ai joint sa sous-division en 1988. Il m'a immédiatement recruté comme rédacteur adjoint de *Techniques d'enquête*. C'était une pratique courante chez M.P. – lorsque j'ai rencontré des personnes avec un bon bagage technique, elles devenaient des rédacteurs adjoints potentiels de la revue. Ceux d'entre nous qui ont été assez chanceux pour occuper un tel poste ont beaucoup appris de l'expérience. Au fil du temps, lorsque M.P. s'est fié à notre jugement, il s'en est remis de plus en plus à notre opinion, par exemple, pour décider du sort d'un article qui avait fait l'objet d'évaluations contradictoires.

M.P., le gestionnaire : son approche envers ses rédacteurs adjoints illustre bien son style de gestion plus général. Il laissait à chacun le soin de faire ses preuves et, sauf de rares exceptions, les capacités de chaque employé se sont développées parallèlement à la confiance que M.P. leur manifestait. Beaucoup de gestionnaires suivent une philosophie de gestion spécifique, quelquefois sautant sur la dernière tendance de gestion, quelle qu'elle soit. M.P. n'était pas ainsi. Gestionnaire intuitif, il avait le don de déceler les futurs « talents » au tout début de leur carrière. Il était aussi un

ménages à Statistique Canada, et il a aidé à convaincre les gestionnaires de Statistique Canada des mérites potentiels de ce concept.

M.P. a eu une influence majeure à Statistique Canada sur la qualité et le calibre de la recherche en méthodes statistiques. Les réalisations du Bureau dans ce domaine ont été reconnues dans le monde entier et on demande maintenant souvent à Statistique Canada de participer à des activités de recherche, par exemple en présentant des articles à des réunions, en participant à des discussions entre experts et en siégeant à différents comités et groupes consultatifs. Un comité de recherche en méthodologie a été créé en 1982-1983 et M.P. en a été le premier président. C'est là qu'il a aidé à élaborer un programme de recherche et un plan stratégique pour la Direction de la méthodologie. Même si le Programme de recherche a évolué au fil du temps, le Programme de recherche en méthodologie est toujours florissant, grâce à la structure et au soutien de gestion que M.P. a aidé à mettre en place.

Tout au long de ma carrière à Statistique Canada, j'ai pu profiter largement de la présence de M.P. Aux réunions de gestion et aux réunions où il représentait la Direction de la méthodologie, il a toujours veillé à ce que nous conservions nos qualités distinctives comme méthodologistes et à ce que les décisions que nous prenions aient un sens pour notre groupe.

Même avec toutes ses réalisations comme statisticien d'enquête, c'était son tempérament que j'admirais le plus. À mon avis, sa compassion désintéressée pour les autres, quel que soit leur niveau de compétence, était sa plus grande force. Je me souviens d'une occasion où nous étions tous les deux en train d'interviewer à Ottawa un candidat hautement qualifié que nous avions fait venir de très loin. Toutefois, après quelques minutes, il était clair, malgré ses qualifications, que cette personne n'était pas faite pour travailler à la Direction de la méthodologie. Bien que le candidat ait fait un voyage spécial à Ottawa pour l'interview, M.P. a pris le temps de mettre la personne à l'aise en discutant avec elle de sujets familiers, même si M.P. reconnaissait aussi qu'elle n'était pas faite pour la Direction.

M.P. a toujours sa faire l'éloge des autres lorsque leurs réalisations étaient dignes de mention. C'est une des nombreuses raisons pour lesquelles beaucoup l'aimaient et pour-quoil il manquait à plusieurs.

Bibliographie

Kasprzyk, D., Duncan, G., Kalton, G. et Singh, M.P. (Ed.) (1989). *Panel Surveys*. New York : John Wiley & Sons, Inc.

Platiek, R., Rao, J.N.K., Samdal, C.E. et Singh, M.P. (Ed.) (1987). *Small Area Statistics : An International Symposium*. New York : John Wiley & Sons, Inc.

statisticien d'enquête exceptionnel et comme personne attentive et aimable, le définissaient dans une classe à part.

J'ai rencontré M.P. Singh pour la première fois à l'été 1970. Je travaillais comme étudiant à la Division d'agriculture de Statistique Canada, où M.P. Singh et J.C. (John) Koop étaient alors méthodologistes. Je partageais un bureau avec Jack Graham, en congé sabbatique de l'Université Carleton. À ce moment-là, Jack m'avait confié combien Statistique Canada était privilégiée d'avoir M.P. et John comme méthodologistes d'enquête, car ils étaient deux des meilleurs statisticiens d'enquête au monde. Des talents si exceptionnels à Statistique Canada m'ont incité à choisir cet endroit pour amorcer ma carrière.

La plupart des gens connaissaient M.P. par les douzaines d'articles qu'il avait publiés, par ses fonctions à la revue *Techniques d'enquête* et par ses interventions aux conférences statistiques. Ses publications comprenaient des articles sur la conception et la révision des enquêtes auprès des ménages, sur l'estimation (dont l'estimation composite et l'estimation par domaine), sur l'estimation des données régionales et sur les ajustements pour la non-réponse. Ses questions et suggestions lors de conférences et de réunions reflétaient la profondeur de sa pensée sur les nombreuses complexités des méthodes d'enquête.

Il a aussi rédigé en collaboration des monographies sur les enquêtes par panel (Kasprzyk *et al.* 1989) et sur les statistiques régionales (Platak *et al.* 1987), et il a écrit un exposé de synthèse sur les *Techniques d'enquête* dans l'*Encyclopedia of Statistical Sciences* (Singh 1988).

Rédacteur en chef de *Techniques d'enquête* depuis sa fondation en 1975, M.P. a supervisé l'évolution de la revue, au début comme l'outil principal de publication des recherches du personnel de Statistique Canada puis comme une revue internationale de pointe à laquelle ont collaboré régulièrement des auteurs de partout dans le monde. La section sur les méthodes de recherche par sondage de l'American Statistical Association puis l'Association internationale des statisticiens d'enquête adoptèrent *Techniques d'enquête* comme publication pour leurs membres. Cela reflète bien les nombreuses années de travail assidu qu'a données M.P. à la revue. On retrouvait sa gentillesse et son attention même dans les commentaires encourageants qu'il faisait lorsqu'il devait écrire une lettre de refus à un auteur! Au fil des ans, M.P. a été un chef de file dans l'adaptation des enquêtes auprès des ménages à l'évolution de la technologie. Il a toujours cherché des moyens d'améliorer les méthodes de collecte de données. Il a orienté Statistique Canada dans cet univers de l'interview face-à-face, de l'interview téléphonique et des méthodes informatiques. Tout dernièrement, il s'appliquait à élaborer des méthodes plus efficaces, notamment en instaurant le concept d'un échantillon-maître pour la conception d'enquêtes auprès des

examen, la qualité était excellente et, sous sa direction, mon travail consistait à faire en sorte que les numéros de la revue à

paraître soient encore meilleurs. Les remises en question faisaient partie de son travail de rédacteur en chef de *Techniques d'enquête*. La revue devait offrir des formulations statistiques mathématiques très bien soutenues, mais elles devaient aussi pouvoir être mises en application. En d'autres mots, les idées devaient être très bonnes, mais aussi éminemment utiles. Et elles l'ont toujours été. Ce n'est pas un mince exploit.

Beaucoup de jeunes professionnels hors pair, dans leur premier article, démontraient uniquement un de ces aspects, habituellement la dimension mathématique de leur sujet. Selon moi, l'objectif que fixait M.P. à ses rédacteurs associés qui recevaient des articles pressurant au moins un de ces aspects était d'aider les auteurs, grâce à l'examen et à nos commentaires, à atteindre le deuxième objectif. C'est tout un

journal que sa vision a créé!

À propos, il m'avait confié que j'avais peut-être tendance à en faire trop dans mon rôle de soutien aux auteurs, mais je pense que, secrètement, il était heureux de mon approche de ne jamais abandonner un article qui pouvait devenir extraordinaire, si on était suffisamment patient. Plusieurs articles dont je me suis occupé ont éprouvé sa patience mais l'ont récompensé en bout de ligne.

M.P. avait une ténacité qui complétait son indépassable bonté. Sa direction ferme et sûre de *Techniques d'enquête* nous obligeait tous à respecter des normes élevées. Même quand sa santé a commencé à décliner, son esprit est toujours demeuré manifeste.

Le mot que j'ai utilisé pour caractériser M.P. à la conférence de l'autonomie dernier était celui de « Mensch ». Ce mot allemand désignant une « personne » peut être familier pour nombre d'entre vous dans son sens yiddish d'un être humain complet ou entier. Mais en réalité, le mot « Mensch » ne peut vraiment pas se traduire. C'est pourquoi je l'ai laissé en yiddish ici (même si je n'ai pas utilisé de caractères hébreux, ce qui aurait été approprié). Il n'y a certainement pas de définition simple qui puisse rendre justice soit au mot, soit à la personne qu'a été M.P.

Il nous manque tous énormément. Il était un bon ami, un homme de famille tendre, ouvert aux idées nouvelles, prudent dans ses conseils pratiques et rigoureux dans sa pensée. Il sera toujours un modèle de ce qu'est un statisticien d'enquête.

Quelques souvenirs de M.P. Singh

David A. Binder

Statistique Canada (retraitée)

Je garde de très bons souvenirs de M.P. Singh, que j'ai pendant de nombreuses années. Ses forces, comme

Les efforts déployés pour faire de l'EPA la base d'autres enquêtes auprès des ménages connaurent un tel succès, qu'un problème de surcharge survint à la fin des années 90. Avec l'ajout des enquêtes longitudinales et des enquêtes sur la santé au programme d'enquêtes régulier, on se préoccupa davantage du fardeau de réponse, on sentait le besoin d'avoir des bases de sondage plus ciblées pour certaines sous-populations. M.P., chercha alors d'autres solutions dont certaines axées sur le registre des adresses élaboré aux fins de recensement. Quelques-unes de ces approches ont été intégrées dans la révision de l'EPA après le recensement de 2001, révision amorcée au moment de son décès; d'autres idées plus ambitieuses pour une nouvelle base de sondage aux enquêtes auprès des ménages sont toujours à l'étude par ses successeurs.

Pendant plus de 30 ans, M.P. a orienté les travaux méthodologique à l'EPA. Ses nombreux articles, souvent rédigés en collaboration avec son personnel, témoignent de son influence permanente sur la conception de cette enquête phare et sur l'évolution qu'il connaît de nombreux jeunes statisticiens au début de leur carrière.

Au cours de cette même période, M.P. a aussi assumé une autre lourde responsabilité soit celle de rédacteur en chef de *Techniques d'enquête*. L'évolution de cette revue, de sa naissance en 1975 jusqu'à son 25^e anniversaire, a été décrite par son fondateur, Richard Platak (1999), qui avait eu la clarté voyante de nommer M.P. comme son premier rédacteur en chef.

Sous le leadership de M.P., la revue a franchi beaucoup d'étapes importantes. En 1982, elle est devenue une publication officielle de Statistique Canada – entièrement bilingue et tarifée. On invita des auteurs de l'extérieur de Statistique Canada; un panel hautement qualifié de rédacteurs associés fut recruté; on presenta des numéros thématiques, qui attirèrent souvent les meilleurs articles d'une récente conférence ou symposium; on institua la rubrique *Dans ce numéro* où le rédacteur en chef présentait une vue d'ensemble du contenu; des numéros spéciaux du 25^e anniversaire parurent en 1999–2000, accompagnés d'un index des volumes 1 à 26. Pendant cette période, on offrit, d'abord à l'Association internationale de statisticiens d'enquêtes puis d'autres organismes statistiques, des abonnements à prix réduit. Plus récemment, des versions électroniques de la revue sont devenues disponibles.

Au cours de cette période, M.P. a tenu la barre, plantant les numéros à venir, à l'affût de travaux intéressants dignes d'être inclus dans la revue, encourageant des auteurs potentiels, recrutant et harcelant les rédacteurs associés au cours du processus d'examen, travaillant avec le personnel de la publication et du marketing de Statistique Canada pour améliorer la revue et en faire la promotion. En tant que membre du Conseil de gestion de la revue de 1987 à 2004,

J'ai été à même de constater et d'admirer son enthousiasme et sa persévérance face à de nombreuses difficultés. C'était pour lui, je crois, une œuvre d'amour.

Ces brèves descriptions de seulement deux des multiples contributions de M.P. à Statistique Canada et à la profession statistique ne peuvent rendre pleinement justice à sa carrière. J'espère qu'elles donnent l'image d'un professionnel sur qui on pouvait toujours compter, qui savait allier une capacité de compréhension profonde et de recherche des méthodes statistiques à une connaissance des contraintes pratiques que comporte l'application des méthodes statistiques aux enquêtes. Son style s'appuyait sur la raison et la persistance, sans éclat et dans l'acceptation des contradictions, auxquelles s'ajoutait une préoccupation innée pour les sentiments des autres. J'ai toujours pris plaisir à travailler avec M.P., et je suis honoré d'être associé à ses réalisations.

Bibliographie

- Drew, D., Singh, M.P., et Choudhry, H. (1982). Évaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active du Canada. *Techniques d'enquête*, 8, 19-52.
- Platak, R., et Singh, M.P. (1976). *Methodology of the Canadian Labour Force Survey*, Statistique Canada, numéro de catalogue 71-526.
- Platak, R. (1999). Techniques d'enquête – 25 ans d'histoire. *Techniques d'enquête*, 25, 123-125.
- Statistique Canada (1990). *Methodology of the Canadian Labour Force Survey 1984-1990*, Statistique Canada, numéro de catalogue 71-526.

À la mémoire de M.P. Singh

Fritz Scheuren
président de
l'American Statistical Association pour 2005

Avec le décès l'été dernier de M.P. Singh, la communauté statistique entière perd un érudit, un homme d'honneur et un homme d'action. C'est en ces termes que j'ai parlé de lui au Symposium sur la méthodologie de Statistique Canada à l'automne 2005.

Toutefois, je serai bref, domant seulement un exemple de ce qui pourrait être dit. D'autres parlant aussi de lui. Ils en diront plus.

Mes souvenirs de M.P. remontent à plus de 20 ans. Je ne me souviens pas exactement à quel moment j'ai rencontré pour la première fois, mais j'ai été un de ses rédacteurs associés à *Techniques d'enquête* pendant au moins tout ce temps.

Il aimait me faire lire des articles sur le couplage d'engagements, quelquefois sur la pondération ou l'estimation, et moins souvent, sur des sujets reliés aux données manquantes. Ses choix étaient toujours pour moi une occasion d'apprendre. De façon générale, après son premier

désigné comme président du Comité de recherche sur la méthodologie. À ce poste jusqu'en 1987, il a mis en place les processus de planification et les critères de déclaration qui, améliorés par ses successeurs, ont régi la gestion de la recherche méthodologique pendant deux décennies. C'est au cours de cette même période que Statistique Canada inaugura les symposiums sur la méthodologie, M.P. jouant un rôle clé dans plusieurs des premiers symposiums (et dans beaucoup d'autres par la suite).

Au cours de sa longue carrière à Statistique Canada, M.P. a participé à une vaste gamme de travaux méthodologiques, mais on associe toujours son nom de façon plus immédiate à deux projets : la conception de l'Enquête sur la population active du Canada (EPA) et la fonction de rédacteur en chef de la revue *Techniques d'enquête*.

L'EPA est la base du programme des enquêtes auprès des ménages de Statistique Canada. Non seulement est-elle la source des estimations mensuelles sur les conditions du marché du travail au Canada, mais sa base de sondage est aussi la base d'échantillonnage de nombreux autres enquêtes auprès des ménages, dont plusieurs enquêtes longitudinales des années 90. Sa conception efficace est donc cruciale à la rentabilité du programme de statistiques sociales du Canada. D'abord inaugurée en 1945, l'EPA a typiquement subi au moins une révision de son échantillon après chaque recensement décennal. M.P. a joint Statistique Canada juste à temps pour la révision majeure qui a suivi le recensement de 1971. Cette révision incluait non seulement le plan de sondage mais aussi le questionnaire, les méthodes de collecte et les systèmes de traitement. Une révision si importante nécessitait une vaste collaboration interdisciplinaire et M.P. a été un intervenant clé dans les aspects méthodologiques de cette révision. Ses articles de cette période traitent surtout de l'optimisation du plan à plusieurs degrés et de la mise à jour de la méthodologie de l'Enquête sur la population active (Platak et Singh 1976).

À la suite de cette révision, les pressions pour la production d'estimations régionales du marché du travail s'accroissent et amènent M.P. à élaborer des méthodes d'estimation de données régionales à partir de l'EPA (Drew, Singh et Choudhry 1982). Au moment de la révision prévue après le recensement de 1981, M.P. était devenu le président du comité chargé de superviser le processus complet de révision. En plus des objectifs habituels d'efficacité de l'échantillon-nage, cette révision avait pour but de produire de meilleures données infraprovinciales et d'améliorer le rôle de l'EPA comme véhicule pour la réalisation d'autres enquêtes auprès des ménages. Naturellement, M.P. a encore été un des auteurs principaux derrière la description du nouveau plan de sondage (Statistique Canada 1990).

donné le numéro de téléphone de M.P. à Ottawa. Alors, quand nous sommes arrivés à Ottawa, j'ai appelé M.P. de l'hôtel où nous restions et, à ma surprise, il est venu me chercher par un matin froid et humide de la fin de septembre et m'a emmené à Statistique Canada. Ce geste chaleureux et amical a égayé ma journée et mon arrivée à Statistique Canada.

M.P. était originaire de la ville ancienne de Bénarès, en Inde, et il semble que certaines des qualités qui ont rendu cette ville célèbre avaient déteint sur lui. Il était doux, gentil, imperturbable, tenace et sage. Beaucoup de gens m'ont dit qu'il n'était jamais trop occupé pour écouter leurs problèmes et qu'il les aidait toujours avec des paroles bienveillantes et des suggestions. Beaucoup de jeunes statisticiens ont profité de ses conseils concernant leurs recherches et leur carrière.

M.P. aimait la musique et la danse indienne classique. Sa famille comptait beaucoup pour lui. Il était le pilier sur lequel s'appuyaient sa femme et ses enfants lorsqu'ils éprouvaient des difficultés. Lors des rencontres sociales, il avait toujours beaucoup de plaisir et était énormément. Et lorsqu'il est tombé gravement malade il y a quelques années, il m'a dit que c'était sa foi en Dieu et en lui-même qui l'avait aidé à se rétablir. On se souviendra longtemps de lui, non seulement comme un statisticien de renom, mais aussi comme d'un brave homme qui s'est lié d'amitié avec beaucoup de gens et qui en a aidé plus d'un.

Une carrière en méthodologie d'enquête

Gordon Brackstone

Statistique Canada (retraité)

Presque toute la carrière de M.P. Singh s'est déroulée dans le secteur de la méthodologie de Statistique Canada. Il a joint l'organisme en 1970, après avoir obtenu un doctorat en échantillonnage de l'Indian Statistical Institute. Au moment de son décès, il était directeur de la Division des méthodes d'enquêtes auprès des ménages à la Direction de la méthodologie. Sa montée dans l'organisme a été constante pluri que vertigineuse : chef de section en 1973, directeur adjoint en 1982 puis directeur en 1994. Cette progression constante reflète bien son approche de la méthodologie d'enquête qui privilégiait le souci du détail dans la recherche et les essais afin d'établir de solides fondations pour la mise en œuvre et les améliorations futures.

Nos carrières à Statistique Canada ont coïncidé, à une année près au début ou à la fin, et elles se sont souvent croisées, particulièrement à partir de 1982. Au début des années 80, lorsque nous avons senti la nécessité d'améliorer l'intégration et la surveillance de la recherche méthodologique à Statistique Canada, j'étais à peu près certain qu'on demanderait à M.P. de diriger ce travail, et il a été dûment

dit que M.P. avait lu dans sa propre main la fin de ses graves problèmes de santé. Avi et moi étions certains de revoir M.P. au travail. Toutefois, c'est une croyance populaire en Inde que ceux qui lisent dans leur propre main ne peuvent prévoir leur avenir avec précision. Malheureusement, cette croyance s'est avérée juste dans ce cas-ci.

M.P. était vraiment un grand ami et il me manquera beaucoup. Il convient que ses cendres aient été répandues dans la fleuve sacré Gange, dans la ville la plus sainte pour les Hindous. Varanasi (aussi appelée Bénarès), où était né M.P. Son âme est au Ciel mais son héritage demeurera avec nous.

M.P. du temps où il était chercheur

T.J. Rao

Indian Statistical Institute, Kolkata

J'ai rencontré M.P. pour la première fois lorsqu'il a assisté

au quatrième cours d'été (avancé) pour statisticiens organisés par la Research and Training School (RTS) de l'Indian Statistical Institute (ISI) en mai et en juin 1964 à l'Université de Kerala située dans la ville de Trivandrum (maintenant appelée Thiruvananthapuram) dans le sud de l'Inde. Ce cours était destiné aux chercheurs et aux professeurs débutants de l'ISI et d'autres universités. M.P. venait de l'Université Banaras Hindu (BHU) où il était un chargé de cours temporaire. Il a obtenu un baccalauréat en statistique de la même université (BHU) et une maîtrise de l'Université de Poona. J'étais parmi les chercheurs qui ont été sélectionnés par l'ISI pour participer à ce cours. Nous n'avons pas beaucoup interacté pendant le cours.

Un peu plus tard, M.P. a reçu une offre d'emploi de la Division de l'échantillonnage du département National Sample Survey (NSS), qui faisait partie de l'ISI à l'époque. Les professeurs D.B. Lahiri, S. Rajaratn et M.N. Murthy dirigeaient déjà alors plusieurs divisions du NSS. En plus de travailler à la conception d'enquêtes par sondage à grande échelle réalisées par le NSS, M.P. passait son temps libre à examiner des problèmes de recherche liés aux enquêtes par sondage. Lahiri et Murthy encourageaient la recherche méthodologique au NSS et ont commencé à organiser une série de séminaires ainsi qu'à diffuser des rapports techniques semblables aux rapports techniques de la RTS de l'ISI. M.P. et moi avons discuté de notre recherche sur les problèmes d'échantillonnage au cours des séminaires organisés par le NSS et la RTS. La majeure partie des travaux de M.P., qu'il a converti en rapports techniques pour la série du NSS, ont été publiés plus tard dans des revues scientifiques réputées. En s'appuyant sur l'expertise qu'il avait acquise au NSS en travaillant sur des enquêtes polyvalentes, il s'est intéressé aux problèmes liés à l'utilisation de données auxiliaires dans des enquêtes par sondage. Ses premiers travaux ont porté sur l'estimation par les méthodes du quotient et du produit. M.P.

Naajamma Chinappa Statistique Canada (retraitée)

M.P. Singh

Singh, M.P. (1967). Ratio cum product method of estimation. *Metrika*, 12, 34-42.
Singh, M.P. (1969). Some aspects of estimation in sampling from finite populations. Thèse de doctorat, soumise à l'Indian Statistical Institute.
Murthy, M.N., et Singh, M.P. (1969). On the concepts of best and admissible estimators in sampling theory. *Sankhyā*, 31, 343-354.

M.P. aimait beaucoup assister à des conférences. Il n'en a jamais manqué une à son alma mater BHU ou au Indian Science Congress. Il a entrepris la rédaction de sa thèse avec beaucoup de sérieux et il aimait discuter avec les professeurs M.N. Murthy, J.N.K. Rao et D. Basu. Il a présenté ses travaux de recherche comme thèse (Singh 1969) pour obtenir un doctorat en philosophie (Ph.D.) de l'Indian Statistical Institute en 1969 sous la direction de M.N. Murthy. Il a quitté le NSS et l'ISI en 1970 pour se joindre à Statistique Canada. Il manque beaucoup à tous les chercheurs qui étaient à l'ISI entre 1965 et 1970 et à ses collègues du NSS.

Il a étudié intelligemment et avec succès le cas de multiples variables auxiliaires dont certaines étaient corrélées positivement et certaines étaient corrélées négativement avec la variable étudiée. Il a utilisé les estimateurs par quotient pour les variables corrélées positivement et les estimateurs par produit pour les variables corrélées négativement et a rédigé le *Ratio cum product estimator* (Singh 1967). Ce document est souvent cité et plusieurs chercheurs, en particulier de l'Inde, ont publié des suppléments. Conjointement avec M.N. Murthy, il a élaboré des concepts intéressants sur l'admissibilité des estimateurs (Murthy et Singh 1969). Au cours de l'année 1968, le professeur J.N.K. Rao a visité l'ISI et nous avons été très chanceux d'interagir avec lui.

Comme beaucoup d'entre vous connaissent M.P., le statisticien, ainsi que ses réalisations en statistique, je vais essayer de vous parler de M.P., l'homme. Je n'avais jamais rencontré M.P. avant mon arrivée au Canada, même si j'avais entendu dire qu'il était le jeune homme qui avait été nommé à mon poste lorsque j'ai démissionné de mon emploi au département National Sample Survey (NSS) de l'Indian Statistical Institute de Kolkata, en Inde. J'ai appris que, lorsque M.N. Murthy (alors chef du secteur de la méthodologie du NSS) m'a envoyé l'ébauche de son livre *Sampling Theory and Methods* pour que je la lise, M.P. est celui qui a lu mes commentaires et en a discuté avec M.N. Murthy. Beaucoup plus tard, lorsque M.N. Murthy a appris que j'avais été embauché par Statistique Canada, il m'a

régression généralisée. Trois communications sur l'estimation composite par régression pour l'EPA, notamment une communication de M.P., Jack Gambino et Brian Kennedy, ont paru dans le numéro de juin 2001 de *Techniques d'enquête*.

Depuis 1976, M.P. s'intéressait aussi vivement à l'estimation des données régionales. Son équipe a fait d'importantes contributions méthodologiques aux estimations des données régionales. M.P. et ses collègues proposèrent des estimateurs synthétiques simples de même qu'un nouvel estimateur appelé l'estimateur dépendant de l'échantillon. Cet estimateur est un estimateur composite simple dont les coefficients de pondération s'appliquent aux tailles d'échantillon réalisées qui sont inférieures aux tailles d'échantillon prévues dans les régions. Les estimateurs dépendants de l'échantillon sont alors devenus assez connus et bon nombre d'organismes dans le monde les ont utilisés. En 1994, M.P. publia dans *Techniques d'enquête*, avec Jack Gambino et Harold Mantel, un document expliquant plusieurs questions pratiques liées à l'estimation des données régionales. L'aine tout particulièrement la section sur les questions de plan de sondage. Elle illustre à merveille le compromis qu'a fait l'EPA en ce qui concerne la répartition de l'échantillon pour satisfaire aux critères de fiabilité tant à l'échelle provinciale qu'infra-provinciale. Un chapitre de mon ouvrage sur l'estimation des données régionales (Rao 2003) est consacré aux questions de planification, lesquelles reposent fortement sur ce document de 1994. M.P. participa activement à l'organisation d'une conférence internationale fort réussie en 1985 sur les estimations des données régionales et il rédigea en collaboration en 1987 chez Wiley un ouvrage intitulé *Small Area Statistics*, qui s'inspire des communications sollicitées à la conférence.

M.P. adorait son travail de rédacteur en chef à *Techniques d'enquête*. Il a maintenu des liens étroits avec son équipe de rédacteurs associés, proposant beaucoup de nouvelles idées dont des exposés thématiques tant sur la théorie que sur la pratique, de même que la série d'articles Wakberg. Les défuntes-causes-organisées chaque année par M.P. aux *Joint Statistical Meetings* ont toujours connu un grand succès auprès des rédacteurs associés! À titre de rédacteur associé à Ottawa et de consultant pour Statistique Canada, j'ai souvent abordé avec M.P. les problèmes auxquels faisait face la revue pendant les 25 dernières années. M.P. a aussi joué un rôle actif à la Société statistique du Canada (SSC) et il a fait la promotion de la théorie de l'échantillonnage aux assemblées générales annuelles de la SSC.

M.P. était un chironomien d'une remarquable précision; en 1999, il m'avait mis en garde au sujet d'éventuels problèmes de santé. De fait un problème de santé imprévu est survenu en 2001 en raison de complications à la suite d'une appendicite. Quelques mois avant son décès, Avi Singh m'a

notre rapport technique de 1969 à l'ISI. Nos résultats démontrent qu'il n'était pas pratique d'utiliser un critère d'échantillonnage basé sur un choix unique. M.P. a aussi démontré que l'application de l'hypéradmissibilité à l'estimation de la variance donnait comme choix unique un « mauvais » estimateur de la variance.

Peu de temps après avoir joint les rangs de Statistique Canada en 1970 à titre de méthodologiste, M.P. a participé activement à la révision de l'EPA qui a débouché sur plusieurs innovations. M.P. a proposé l'utilisation systématique de l'échantillonnage PPT sans remplacement avec randomisation initiale pour la sélection des unités primaires à partir des unités non autoréprésentatives (UNAR) et la méthode des groupes aléatoires avec une unité primaire provenant de chaque groupe aléatoire prélevé par échantillonnage PPT auprès des unités autoréprésentatives (UAR). Dans les années 60, j'avais étudié la théorie de ces méthodes du point de vue de leur efficacité et de l'estimation de la variance. De son côté, M.P. a reconnu leurs avantages pratiques dans le contexte de l'EPA. L'échantillonnage systématique avec PPT et la méthode des groupes aléatoires autorisaient une expansion de l'échantillon de même qu'un renouvellement plus facile des unités primaires d'échantillonnage dans le temps, tandis que la méthode des groupes aléatoires permettait d'adapter la méthode ingénieuse de Keyfitz pour modifier les mesures de taille périmées dans chaque groupe aléatoire. Il publia une communication dans *Mechika* (1975) conjointement avec Dick Platek sur la mise à jour des mesures de taille. Sous l'habile direction de M.P., le groupe de l'EPA apporta plusieurs améliorations méthodologiques à l'efficacité du plan et à l'estimation. Étant donné que M.P. s'était intéressé à l'utilisation efficace des renseignements auxiliaires, l'EPA adopta l'estimation par régression généralisée pour tenir compte de plusieurs variables pour la stratification a posteriori. Le groupe de l'EPA a aussi été le premier à admettre les mérites de l'estimation de la variance par ré-échantillonnage, et on adopta la méthode du jackknife pour l'estimation de la variance. Plus récemment, sous la direction de M.P., on a instauré dans l'EPA l'estimation composite par régression en s'inspirant d'une méthode proposée par Wayne Fuller et moi-même, qui est utile à la fois pour l'estimation du changement et du niveau. Cette méthode de même qu'une méthode antérieure d'Avi Singh s'harmonisent bien avec le système actuel d'estimations de l'EPA qui repose sur la

A la mémoire de M.P. Singh

Introduction

Don Royce
Statistique Canada

En août 2005, le monde de la méthodologie d'enquête a perdu l'une de ses figures dominantes avec le décès de M. M.P. Singh à l'âge de 63 ans, quelques mois avant sa retraite planifiée. M.P. et moi avions discuté brièvement de sa retraite à venir, mais il était clair pour nous deux qu'il continuerait d'être le rédacteur en chef de *Techniques d'enquête* même après son départ de Statistique Canada. *Techniques d'enquête* faisait partie de sa vie et j'étais très heureux de lui offrir l'occasion de travailler à temps partiel à partir de sa résidence familiale à Toronto, ce qui lui permettait de continuer à s'occuper de la revue qu'il avait dirigée depuis plus de 30 ans. Malheureusement, cela n'a jamais eu lieu.

Dans la série d'articles qui suivent, nombre de collègues et amis (les deux sont synonymes) les plus proches de M.P. se rappellent de lui comme statisticien, rédacteur en chef, collaborateur, leader et être humain. Je suis profondément reconnaissant à Eric Rancourt de Statistique Canada d'avoir proposé cette série d'articles, et à tous les auteurs qui, par leur temps et leur talent, ont partagé leurs souvenirs de M.P. Singh. Même si les mots ne peuvent jamais traduire complètement l'essence d'une personne, les articles qui suivent décrivent merveilleusement bien la vie de M.P. Singh et nous rappellent l'héritage qu'il laisse à tous ceux qui ont eu la chance de le connaître. Nous espérons que M.P. en aurait été heureux.

Quelques souvenirs

J.N.K. Rao
Université Carleton, Ottawa

Ma première rencontre avec Mangala Prasad Singh (ami-lorsque j'étais professeur invité à l'Indian Statistical Institute à Calcutta, M.P. poursuivait son doctorat à l'ISI sous la supervision de M.N. Murthy. Pendant son doctorat, il a aussi travaillé au National Sample Survey (NSS) en Inde. Le NSS était situé sur le campus de l'ISI et M.P. y a travaillé sous la direction de statisticiens d'enquête renommés au NSS et à l'ISI, dont P.C. Mahalanobis, D.B. Lahiri et M.N. Murthy. Il

a ainsi reçu une solide formation en conception et en théorie d'enquêtes par sondage. M.P. a fait un bon usage de cette solide formation pendant toute sa carrière en appliquant les principes d'une conception efficace sous réserve des questions de coût et d'opération, et en insistant sur l'importance d'une théorie robuste avant de mettre en œuvre de nouveaux plans d'enquête ou de réviser des enquêtes permanentes comme l'Enquête sur la population active du Canada (EPA). Une grande partie de la thèse de M.P. Singh portait sur l'utilisation efficace des renseignements auxiliaires. Il a étudié deux variables auxiliaires, une en corrélation positive et l'autre en corrélation négative avec la variable d'intérêt, et il a conçu des estimateurs de totaux ratio-produit (ratio-cum-product). Murthy (1967), dans son ouvrage bien connu sur l'échantillonnage, a consacré une section aux estimateurs ratio-produit. M.P. a publié plusieurs documents sur l'utilisation efficace des renseignements auxiliaires développée dans sa thèse : les estimateurs ratio-produit (*Metrika* 1967; *Sanhitya* 1969), l'estimation multidimensionnelle de produits (*Journal of the Indian Society of Agricultural Statistics* 1967) et l'échantillonnage systématique dans l'estimation des ratios relative des stratégies d'échantillonnage à deux phases dans un modèle de superpopulation. La première phase consistait en un échantillonnage aléatoire simple servant à recueillir des données sur une variable auxiliaire x utilisée dans la deuxième phase pour choisir un échantillon PPT sans remplacement et collecter des données sur la variable d'intérêt y . Au moment de ma visite à l'ISI, M.P. explorait aussi des questions d'inférence en échantillonnage et il faisait face à des problèmes techniques pour démontrer l'admissibilité de certains estimateurs. En effet, un estimateur est admissible dans une classe d'estimateurs non biaisés si aucun autre estimateur de la classe n'est uniformément plus efficace. Malheureusement, le critère d'admissibilité n'est pas suffisamment sélectif et, pour cette raison, la documentation statistique proposait comme choix unique d'autres critères liés à l'admissibilité. Comme je m'intéressais aussi aux questions d'inférence à ce moment-là, nous avons commencé à travailler ensemble sur l'admissibilité des estimateurs. Il a intégré à sa thèse de doctorat le résultat de notre travail. À la fin, nous avons publié un document reposant sur ce travail dans l'*Australian Journal of Statistics* (1973) qui était fondé

Dans l'article de Th  berge, on propose une nouvelle approche pour r  partir l'  chantillon de la contre-v  rification des dossiers (CVD) de 2006 qui vise    mesurer le sous-d  nombrement du recensement et une partie du sur-d  nombrement. Les estimations de la CVD sont utilis  es conjointement avec les chiffres du recensement pour produire des estimations d  mographiques, lesquelles servent      tablir les paiements de p  r  quation du gouvernement f  d  ral canadien aux provinces. L'approche propos  e permet d'  tablir une r  partition qui fournit un   quilibre entre quatre objectifs. Elle consiste d'abord    calculer une r  partition distincte pour chaque objectif. On prend ensuite pour chaque province la taille d'  chantillon maximale sur chacune des r  partitions. La r  partition infraprovinciale de l'  chantillon de la CVD est obtenue en utilisant la technique du calage pour effectuer un lissage de param  tres d  finis au niveau des strates.

Dans son article, Longford discute de la fa  on de concevoir une enq  te lorsque l'on doit produire des estimations pour plusieurs petits domaines, pour lesquels les priorit  s varient   ventuellement, par minimisation d'une somme pond  r  e des variances esp  r  es. Il commence par d  velopper ses id  es dans le contexte de l'estimation directe, puis les   tend    l'estimation composite qui combine l'estimateur direct    un estimateur synth  tique. Pour illustrer les m  thodes, il pr  sente les r  sultats, sous diverses hypoth  ses, de la r  partition de l'  chantillon d'une enq  te aupr  s des m  nages entre les divers cantons suisses.

Vous et Chapman proposent une approche hi  rarchique bay  sienne de l'estimation pour petits domaines lorsque les erreurs d'  chantillonnage des estimateurs directs sont estim  es. Ils d  montrent leur approche en produisant des estimations pour petits domaines    partir de deux ensembles de donn  es et   tudient sa sensibilit   aux hypoth  ses de mod  lisation.

Khoshgooyanfar et Monazzaab comparent des m  thodes d'estimation pour petits domaines bas  es sur un estimateur synth  tique, un estimateur composite et un estimateur empirique bay  sien en vue de produire des estimations intercensitaires des taux provinciaux de ch  mage en Iran. Ils constatent que l'estimateur composite et l'estimateur empirique bay  sien produisent l'un et l'autre des r  sultats satisfaisants.

La br  ve note de Gabler, H  der et Lym, qui conclut ce num  ro, constitue une extension int  ressante de l'article publi   ant  rieurement par Gabler, H  der et Lahiri dans *Techniques d'enq  te* (1999). Elle offre une solution pratique au probl  me de la d  termination de l'effet de plan lorsque des   chantillons diff  rents sont utilis  s pour diff  rents domaines exclusifs.

Enfin, nous tenons    souligner que *Techniques d'enq  te* est maintenant disponible en ligne dans un format PDF enti  rement interrogeable. Tous les articles publi  s dans la revue peuvent d  sormais   tre consult  s gratuitement en direct sur le site Web de Statistique Canada des leur diffusion. Nous pr  voyons   galement inclure les num  ros ant  rieurs. Tous les articles parus dans les sept derniers num  ros sont d  j   mis en ligne et les travaux se poursuivent en vue d'ajouter ceux qui ont   t   publi  s au cours des dix ann  es ant  rieures. Une version imprim  e de la revue continue d'  tre produite pour les abandonn  s. Les anciens num  ros peuvent   tre obtenus sur demande en version imprim  e ou en format PDF scann  . La revue peut   tre consult  e sur le site Web de Statistique Canada    l'adresse <http://www.statcan.ca/bsolc/francais/bsolc?catno=12-001-X>.

Harold Mantel, R  dacteur en chef d  l  gu  

Dans ce numéro

Ce numéro de la revue *Techniques d'enquête* débute par un article spécial à la mémoire de M. P. Singh, rédacteur en chef fondateur, qui a dirigé la revue pendant 30 ans et en a fait la source internationalement reconnue d'information sur les derniers progrès concernant les techniques d'enquête et les méthodes de production de statistiques officielles qu'elle est aujourd'hui. Un grand nombre de collègues et amis proches qu'a comptés M. P. au fil des ans y partagent leurs souvenirs et évoquent sa carrière et ses contributions.

Dans le premier article ordinaire du présent numéro, Thompson discute de l'utilisation des plans de sondage à marche aléatoire pour l'échantillonnage d'une population résauée. Il montre comment cette approche peut mener à des échantillons en réseau où les probabilités d'inclusion peuvent être estimées indépendamment de la façon dont l'échantillon initial de noeuds est choisi, ce qui donne des méthodes valides d'inférence fondée sur le plan de sondage. La sélection préférentielle de certains types de noeuds ou caractéristiques des graphes est possible grâce au choix du mécanisme de marche aléatoire. Il décrit des plans de sondage à marche aléatoire uniforme ainsi que ciblé, et présente certains exemples.

Durant et Skinner examinent le recours à l'imputation et à la pondération pour corriger l'erreur de mesure dans l'estimation d'une fonction de répartition. Ils étudient diverses méthodes d'imputation par le plus proche voisin et d'imputation hot-deck, ainsi que la pondération par le score de propension à répondre sous divers modèles de réponse. Ils discutent des propriétés théoriques de ces méthodes et les comparent au moyen de simulations afin d'estimer la distribution de la rémunération horaire au Royaume-Uni d'après des données provenant de l'Enquête sur la population active. Ils concluent qu'une approche fondée sur l'imputation fractionnaire semble être celle qui, dans l'ensemble, est la plus efficace et la plus robuste.

Harms et Duchesne étudient le problème de l'estimation des quantiles en utilisant des données d'enquête. Ils calculent une estimation interpolée d'une fonction de répartition sur des quantiles donnés d'une variable auxiliaire, puis inversent l'estimateur interpolé calculé résultant de la fonction de répartition de la variable d'intérêt. Enfin, ils réalisent une étude par simulation afin de comparer leur approche à d'autres méthodes.

Dans leur article, Haziza et Rao proposent une nouvelle méthode d'imputation par la régression avec utilisation des probabilités de réponse. Cette nouvelle méthode mène à des estimateurs valides sous l'approche du modèle de non-réponse ou sous celle du modèle d'imputation. Sous la première approche, le mécanisme de réponse est modélisé paramétriquement et n'est pas limité au modèle de non-réponse uniforme, tandis que sous la seconde, les variables d'intérêt sont modélisées et la non-réponse est considérée comme étant ignorable. Les auteurs fournissent aussi des estimateurs de la variance sous leur méthode d'imputation. Ils présentent, pour l'estimation ponctuelle ainsi que l'estimation de la variance, des résultats de simulation qui témoignent des bonnes propriétés de la méthode proposée d'imputation par la régression.

L'article de Zanutto et Zaslavsky traite du problème de l'estimation dans le cas du recensement décennal de la population des États-Unis sous échantillonnage pour le suivi des non-répondants. Au lieu d'essayer d'obtenir l'information auprès de tous les non-répondants, un échantillon est tiré pour le suivi, ce qui pose un problème d'estimation pour petits domaines. La stratégie proposée consiste à prédire le nombre de ménages non répondants dans diverses catégories au moyen d'un modèle hiérarchique logarithmique, puis à imputer des renseignements détaillés sur les personnes et les ménages selon la méthode d'imputation par donneur. L'idée, à la première étape, est de modéliser les caractéristiques du ménage en utilisant des covariables peu détaillées à des niveaux détaillés de géographie et des covariables plus détaillées à des niveaux plus élevés d'aggrégation géographique. Une étude par simulation indique que les propriétés du modèle proposé se comparent favorablement à celles d'autres modèles.

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 32, numéro 1, juin 2006

Table des matières

Dans ce numéro..... 1

À la mémoire de M.P. Singh..... 3

Articles Réguliers

Steven K. Thompson
Plans de sondage à marche aléatoire ciblée..... 11

Gabriele B. Durrant et Chris Skinner
Utilisation de méthodes de traitement des données manquantes pour corriger l'erreur de mesure
dans une fonction de distribution..... 27

Torsten Harns et Pierre Duchesne
De l'estimation des quantiles par calage..... 41

David Haziza et Jon N.K. Rao
Une approche fondée sur un modèle de non-réponse à des fins d'inférence en présence d'imputation
pour des données d'enquête manquantes..... 59

Elaine L. Zanutto et Alan M. Zaslavsky
Un modèle d'estimation et d'imputation des ménages du recensement non-répondants
sous échantillonnage pour le suivi des cas de non-réponse..... 73

Alain Théberge
Répartition de l'échantillon de la contre-vérification des dossiers de 2006..... 87

Nicholas Tibor Longford
Calcul de la taille de l'échantillon pour l'estimation pour petits domaines..... 97

Yong You et Beatrice Chapman
Estimation pour petits domaines au moyen de modèles régionaux et d'estimations
des variances d'échantillonnage..... 107

Ali-Reza Khoshgooyanfar et Mohammad Taheri Monazzah
Une stratégie rentable d'estimation du chômage au niveau provincial : Une approche
d'estimation pour petits domaines..... 115

Communications brèves

Siegfried Gabler, Sabine Häder et Peter Lynn
Effets de plan pour les échantillons à plans de sondage multiples..... 127

Dédiée à la famille de M.P. Singh : Son épouse, Savitri, et ses enfants, Mala, Mamta et Rahul

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

D. Royce

Anciens présidents G. J. Brackstone

R. Platek

Membres J. Gambino

R. Jones

J. Kovar

H. Mantel

E. Rancourt

COMITÉ DE RÉDACTION

Rédacteur en chef

J. Kovar, *Statistique Canada*

Ancien rédacteur en chef M. P. Singh

Rédacteurs associés

D. A. Binder, *Statistique Canada*

J. M. Brick, *Westat Inc.*

P. Cantwell, *U.S. Bureau of the Census*

J. T. Ethinge, *U.S. Bureau of Labor Statistics*

W. A. Fuller, *Iowa State University*

M. A. Hidiroglou, *Office for National Statistics*

D. Judkins, *Westat Inc.*

P. Kott, *National Agricultural Statistics Service*

P. Lahiri, *JPSM, University of Maryland*

P. Lavallée, *Statistique Canada*

G. Nathan, *Hebrew University*

D. Pfeffermann, *Hebrew University*

N. G. N. Prasad, *University of Alberta*

J. N. K. Rao, *Carleton University*

T. J. Rao, *Indian Statistical Institute*

Rédacteurs adjoints

J.-F. Beaumont, P. Dick, D. Haziza, Z. Patak, S. Rubin-Bleuer et W. Yung, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préférablement en Word au rédacteur en chef, rtc@statcan.ca, Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue.

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente cadastrales. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada: États-Unis 12 \$ CA (6 \$ × 2 exemplaires); autres pays, 30 \$ CA (15 \$ × 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 150 Promenade du Pré Tunney, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada, l'Association Internationale pour la Statistique Officielle et l'Association des statisticiens et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.ca.

Techniques d'enquête



Une revue
éditée
par Statistique Canada
Juin 2006 • Volume 32 • Numéro 1

Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, 2006

Tous droits réservés. Le produit ne peut
être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence.

Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnel, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication de résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada, K1A 0T6.

Juillet 2006

N° 12-001-XPB au catalogue

Périodicité : semestrielle

ISSN 0714-0045

Ottawa

Canada

Statistique
Canada
Statistics
Canada





Statistique
Canada
Statistics
Canada

Canada

Numéro 1

•

Volume 32

•

Juin 2006

Une revue
éditée
par Statistique Canada

N° 12-001-XPB au catalogue

Techniques d'enquête



12-001



Survey Methodology

Catalogue No. 12-001-XPB

A journal
published by
Statistics Canada

December 2006

•

Volume 32

•

Number 2



Statistics
Canada

Statistique
Canada

Canada



Survey Methodology



A journal
published by
Statistics Canada

December 2006 • Volume 32 • Number 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2006

All rights reserved. This product cannot be
reproduced and/or transmitted to any person or organization outside of the licensee's organization.

Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes. This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows: Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s).

Otherwise, users shall seek prior written permission of Licensing Services, Client Services Division,
Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

December 2006

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman D. Royce

Past Chairmen G.J. Brackstone
R. Platek

Members J. Gambino
R. Jones
J. Kovar
H. Mantel
E. Rancourt

EDITORIAL BOARD

Editor J. Kovar, *Statistics Canada*
Deputy Editor H. Mantel, *Statistics Canada*

Past Editor M.P. Singh

Associate Editors

D.A. Binder, *Statistics Canada*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidioglou, *Office for National Statistics*
D. Judkins, *Westat Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
G. Nathan, *Hebrew University*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*

J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
R. Valliant, *JPSM, University of Michigan*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, D. Haziza, Z. Patak, S. Rubin-Bleuer and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$30 (\$15 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.ca.

Survey Methodology
A journal Published by Statistics Canada
Volume 32, Number 2, December 2006

Contents

In This Issue	121
 Waksberg Invited Paper Series	
Alastair Scott Population-Based Case Control Studies	123
 Regular Papers	
Phillip S. Kott Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors	133
Jerome P. Reiter, Trivellore E. Raghunathan and Satkartar K. Kinney The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data	143
Fumio Funaoka, Hiroshi Saigo, Randy R. Sitter and Tsutom Toida Bernoulli Bootstrap for Stratified Multistage Sampling	151
Marcin Kozak and Med Ram Verma Geometric Versus Optimization Approach to Stratification: A Comparison of Efficiency	157
Jean-Claude Deville and Pierre Lavallée Indirect Sampling: The Foundations of the Generalized Weight Share Method	165
Jean-Claude Deville and Myriam Maumy-Bertrand Extension of the Indirect Sampling Method and its Application to Tourism	177
Martín H. Félix-Medina and Pedro E. Monjardin Combining Link-Tracing Sampling and Cluster Sampling to Estimate the Size of Hidden Populations: A Bayesian-Assisted Approach	187
Alan H. Dorfman, Janice Lent, Sylvia G. Leaver and Edward Wegman On Sample Survey Designs for Consumer Price Indexes	197
Neal Thomas, Trivellore E. Raghunathan, Nathaniel Schenker, Myron J. Katzoff and Clifford L. Johnson An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey	217
Acknowledgements	233

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.



In This Issue

This issue of *Survey Methodology* opens with the sixth paper in the annual invited paper series in honour of Joseph Waksberg. It is with sadness that we note the passing of Joseph Waksberg in January of 2005. A short biography of Joseph Waksberg was given in the June 2001 issue of the journal, along with the first paper in the series. For more information about the life and work of Joseph Waksberg, see the Statistical Science article (Vol. 15, No 3) "A Conversation with Joseph Waksberg," by David Morganstein and David Marker available at <http://projecteuclid.org/Dienst/UI/1.0/Home>. I would like to thank the members of the selection committee – David Bellhouse, chair, Gordon Brackstone, Sharon Lohr and Wayne Fuller – for having selected Alastair Scott as the author of this year's Waksberg Award paper.

In his paper entitled "Population-Based Case Control Studies", Scott discusses the analysis of case control studies in which the controls are obtained from a complex sample survey. Using the example of logistic regression, he shows how the survey weighted estimates can be quite inefficient because of the relatively small weight given to the cases. Drawing on an analogy with maximum likelihood estimation, he then proposes a simple, much more efficient alternative that is, however, biased for the intercept term. Efficiency and robustness properties are illustrated through examples. Finally, he briefly discusses the problem of case-control family studies.

Kott considers the use of weight calibration to correct for nonresponse and coverage errors. He gives a general description of calibration estimation, and extends Estavao and Särndal's functional form approach to general calibration. He then discusses properties of this calibration method to correct for unit nonresponse and coverage errors under a quasi-randomization model. He concludes with an empirical example and discussion of some issues.

Reiter, Raghunathan and Kinney investigate through a simulation study the effect of ignoring sampling design variables when building imputation models in a multiple imputation context. They show that potential biases can be reduced by controlling for these design variables in the imputation model, either through a fixed-effect or mixed-effect model. They conclude that a useful prescription for imputers is to include as predictors all variables that are related to the variables being imputed, particularly sampling design variables, so as to make the usual assumption of ignorable non-response satisfied.

The article by Funaoka, Saigo, Sitter and Toida investigates the use of bootstrap variance estimators in stratified multi-stage sampling where the sampling fractions are large. They propose a Bernoulli-type bootstrap that provides consistent bootstrap variance estimates when simple random sampling without replacement is used at each stage. The proposed method is simple to implement and can be extended to any number of stages without much complication. The method is illustrated through a limited simulation study and using data from the 1997 Japanese National Survey of Prices.

In the Kozak and Verma paper, the geometric approach to stratification proposed by Gunning and Horgan (2004) is compared with two optimization approaches; the Lavallée-Hidiroglou algorithm (Lavallée and Hidiroglou 1988) and an optimization algorithm proposed by Kozak (2004). Using five artificial populations of various sizes, the three methods are compared under two scenarios; comparison of the resulting CV under a fixed sample size and comparison of the resulting sample sizes under a fixed level of precision.

Déville and Lavallée present general theoretical foundations for the weight share method in indirect sampling. They define the important concept of a link matrix in indirect sampling, which specifies how the elements of the sampled population are linked to the target population and gives weights to these links that permit unbiased estimation. They discuss important properties of the link matrix, and derive necessary and sufficient conditions for an optimal link matrix to exist. The theory is illustrated with some interesting examples.

Déville and Mauny-Bertrand study the determination of a sampling design and an estimation method for a tourist survey. The main issue that this type of survey has to address is the absence of a sampling frame that can be used to directly reach tourists. To get around this problem, authors suggest to sample services for tourists. This is thus a situation of indirect sampling for which the generalized weight-share method is used to obtain estimates of parameters of interest. Some extensions to the method become necessary. The authors focus more specifically on one of them and describe it in greater detail.

Félix-Medina and Monjardin consider a variant of link-tracing sampling. They use a Bayesian approach to construct estimators of population size, however in order to make inferences about the population size that are robust to erroneous specification of the assumed model, the authors make inferences under the frequentist design-based approach. Based on the results of the simulation study, the proposed estimators perform better than the maximum likelihood estimators that are currently used.

The paper by Dorfman, Lent, Leaver, and Wegman presents a comparison of the Consumer Price Index design methodologies of the United Kingdom and the United States employing the same “scanner” data. They conclude that in the population studied, the UK approach, which involves tighter stratification and, more importantly, more restrictive judgment sampling within strata than the probability sampling of the US approach, does better in estimating a target superlative index. This is shown to be the case, whichever low level price index estimator (the ratio of averages, the geometric mean, or the average of ratios) is employed.

In their paper, Thomas, Raghunathan, Schenker, Katzoff and Johnson use multiple imputation to analyze data with missing values caused by a matrix sampling design. In matrix sampling, only a subset of questions is administered to each respondent in order to reduce respondent burden. The authors develop a method for creating matrix sampling forms, each form containing a subset of questions to be administered to randomly selected respondents. The method is designed so that each form includes questions that are predictive of the excluded questions in order to recover some of the information about the latter. The proposed method and multiple imputation are evaluated using data from the National Health and Nutrition Examination Survey.

Harold Mantel, Deputy Editor

Population-Based Case Control Studies

Alastair Scott¹

Abstract

We discuss methods for the analysis of case-control studies in which the controls are drawn using a complex sample survey. The most straightforward method is the standard survey approach based on weighted versions of population estimating equations. We also look at more efficient methods and compare their robustness to model mis-specification in simple cases. Case-control family studies, where the within-cluster structure is of interest in its own right, are also discussed briefly.

Key Words: Case-control studies; Response-selective sampling; Retrospective sampling; Weighting.

1. Introduction

The case-control study, in which separate samples are drawn from 'cases' (people with a disease of interest, say) and from 'controls' (people without the disease), is one of the most common designs in health research. In fact, Breslow (1996) has described such studies as "the backbone of epidemiology". We shall concentrate on biostatistical applications, but the basic design is an efficient sampling strategy whenever cases are rare and examples are common in many other fields as well (business, social science, ecology, market research, for example). In particular, there has been a parallel development of much of the theory in the econometric literature on choice-based sampling (see Manski and McFadden 1981, Cosslett 1981 for example).

There are two fundamentally different types of case-control study: (set-)matched studies, in which each case is matched with one or more controls, and unmatched studies, in which the case and control samples are drawn independently, although there may be loose "frequency matching", with the control sample allocated across strata defined by basic demographic variables in such a way that the distribution of these variables in the control sample is similar to their expected distribution in the case sample. We are only concerned with unmatched studies here and, more specifically, only with the restricted class of population-based studies in which the controls (and occasionally the cases as well) are selected using standard survey sampling techniques.

An excellent introduction to the strengths and potential pitfalls of case-control sampling is given by Breslow (1996, 2004). One of the most important and difficult challenges confronting anyone designing such a study is to ensure that controls really are drawn from the same population, using the same protocols, as the cases. In the words of Miettinen (1985), cases and controls "should be representative of the same base experience". Failure to ensure this adequately in some early examples led to case-control sampling being

regarded with some suspicion by many researchers. A comprehensive discussion on the principles that should govern the selection of controls is given in Wacholder, McLaughlin, Silverman and Mandel (1991). Since the essence of survey sampling lies in methods for drawing representative samples from a target population, it became natural at some stage to think about using survey methods for obtaining controls. Increasingly over the last 25 years or so, the controls (and occasionally the cases as well) are being drawn using complex stratified multi-stage designs. A good history of this development can be found in Chapter 9 of Korn and Graubard (1999).

The analysis of such studies is a particularly appropriate topic for this paper since Joe Waksberg himself was one of the principal drivers behind the adoption of survey methods (and random digit dialing, in particular) for obtaining controls (see, for example, Waksberg 1998 and DiGaetano and Waksberg 2002).

2. Examples

We start with two examples to illustrate the sort of problem that we want to handle. The first example is typical of the large scale studies conducted by the National Cancer Institute whose personnel have been responsible for much of the development of the area. Joe Waksberg, along with his colleagues at Westat, had a strong influence on the sampling methods used for these studies (see Hartge, Brinton, Rosenthal, Cahill, Hoover and Waksberg 1984, who also gives a description of a number of other similar studies) so it is a natural place to start.

Example 1.

In 1977-78, the National Cancer Institute and the US Environmental Protection Agency conducted a population-based case-control study to examine the effects of ultraviolet radiation on non-melanoma skin cancer over a one-year period (Hartge, Brinton, Rosenthal, Cahill, Hoover and

1. Alastair Scott, Department of Statistics, University of Auckland, Auckland 1, New Zealand. E-mail: a.scott@auckland.ac.nz.

Waksberg 1984, Fears and Gail 2000). The study was conducted at eight geographic locations with varying solar ultraviolet intensities. Samples of non-melanoma skin cancer patients aged 20 to 74 and samples of general population controls from each region were interviewed by telephone to obtain information on risk factors. At each location, a simple random sample of 450 patients and an additional sample of 50 patients in the 20–49 age group were selected for contact. For the controls, 500 households were sampled at each location using Mitofsky-Waksberg random-digit dialing (Waksberg 1978). An attempt was made to interview all adults aged 65–74 as well as a randomly selected individual of each sex aged 20 to 64. In addition, a second Mitofsky-Waksberg sample of between 500 to 2,100 households was taken and information gathered on all adults aged 65 to 74. This resulted in samples of approximately 3,000 cases and 8,000 controls, with the sampling rate for cases being roughly 300 times the rate for controls, depending on age.

The second example is important to me personally since it first introduced Chris Wild and myself to the area.

Example 2.

The Auckland Meningitis Study was commissioned by the NZ Ministry of Health and Health Research Council to study risk factors for meningitis in young children which was reaching epidemic proportions in Auckland at that time (see Baker, McNicholas, Garrett, Jones, Stewart, Koberstein and Lennon 2000). The target population was all children under the age of nine in the Auckland region in 1997–2000.

All cases of meningitis in the target age group over the three year duration of the study were included in the study, resulting in about 250 cases. A similar number of controls was drawn from the remaining children in the study population using a complex multi-stage design. At the first stage of sampling, 300 census mesh blocks (each containing roughly 70 households) were drawn with probabilities proportional to the number of houses in the block. At the second stage, a systematic sample of 20 households was selected from each chosen mesh block and children from these households were selected for the study with varying probabilities that depended on age and ethnicity and were chosen to match the expected frequencies among the cases. Selection probabilities are shown in the table below: (PI means Pacific Islander) Cluster sample sizes varied from one to six and a total of approximately 250 controls was achieved. This corresponds to a sampling fraction of about 1 in 400 on average, so that cases are sampled at a rate that is 400 times that for controls here.

These two studies are fairly typical of the sort of study that we want to discuss. They also illustrate the two main sampling methods used, namely random digit dialing and

area sampling. A lively discussion of the relative merits of these two strategies are given in Brogan, Denniston, Liff, Flagg, Coates and Brinton (2001) and DiGaetano and Waksberg (2002).

Table 1
Selection Probabilities

AGE	MAORI	PACIFIC ISLANDER	OTHER
≤ 1 year	0.29	0.70	0.10
≤ 3 years	0.15	0.50	0.07
≤ 5 years	0.15	0.31	0.04
≤ 8 years	0.15	0.17	0.04

3. General Set-Up

Suppose that we have a binary response variable, Y , with $Y = 1$ denoting a case and $Y = 0$ denoting a control, and a vector of potential explanatory variables, \mathbf{x} . We assume that the value of Y is known for all N units in some target population but that at least some components of \mathbf{x} are unknown. We stratify the population into cases and controls, draw a sample from each stratum based on the variables that we know for all units, and measure the values of the missing covariates for the sampled units (in practice, the control sample is often drawn from the whole population, rather than the units with $Y = 0$. If the proportion of cases is small, the difference will be negligible. Otherwise it is simple to adapt the results below to this variant – for a rigorous development, see Lee, Scott and Wild 2006). Typically, we then want to use the sample data to fit a binary regression model for the marginal probability of a being a case as a function of the covariates. The model used is almost always logistic with

$$\begin{aligned} \text{logit} \{P(Y = 1 \mid \mathbf{x})\} &= \text{log} \left(\frac{P(Y = 1 \mid \mathbf{x})}{P(Y = 0 \mid \mathbf{x})} \right) \\ &= \beta_0 + \mathbf{x}^T \beta_1 \end{aligned} \tag{1}$$

say, where β_0 and β_1 are unknown parameters, and we shall assume model (1) throughout the paper. Extensions to more general regression models are straightforward in principle (see Scott and Wild 2001b) but the resulting expressions are somewhat clumsier than those for the logistic model.

How should we go about fitting the model (1) given sample data? Efficient methods are straightforward with simple or stratified random sampling, but we are interested in more complex sampling procedures here. Very often the complex sampling is simply ignored. Potentially, this could lead to all the usual problems that arise from ignoring sampling design structure. Varying selection probabilities can distort the mean structure and estimates produced by standard programs may be inconsistent. Intra-cluster

correlation can reduce the effective sample size so that routinely-produced standard errors are too small, confidence intervals are too short, p - values too low, and so on. A simple strategy that has been adopted by some researchers to minimize the effect is to keep the numbers of subjects in each cluster small (see Graubard, Fears and Gail 1989, for example). This reduces the design effect and hence the impact of clustering, but it can be a very expensive remedy. We look at some possible ways of coping with standard, more cost-effective, sampling schemes in the next few sections.

4. Survey Weighted Approach

One obvious possibility is to use the standard weighted estimating equation approach embodied in most modern packages for analyzing survey data (see Binder 1983). Suppose first that we had data from the whole finite population. If we assume this finite population is drawn from a superpopulation in which the conditional logistic model (1) holds, then we could estimate β by solving the whole-population or census estimating equations

$$S(\beta) = \sum_i^N \mathbf{x}_i (y_i - p_i(\mathbf{x}_i; \beta)) = 0, \quad (2)$$

where $p_i(\mathbf{x}; \beta) = e^{\beta_0 + \mathbf{x}^T \beta_1} / (1 + e^{\beta_0 + \mathbf{x}^T \beta_1})$. (These are the likelihood equations if population units are assumed to be sampled independently from a superpopulation but the resulting estimators are consistent under much more realistic population structures as long as model (1) holds marginally – see Rao, Scott and Skinner 1998 for more discussion.)

Now, for any fixed value of β , $S(\beta)$ in equation (2) is just a vector of population totals. This means that we can estimate it from the sample, say by

$$\hat{S}(\beta) = \sum_{\text{sample}} w_i \mathbf{x}_i (y_i - p_i(\mathbf{x}_i; \beta)), \quad (3)$$

where w_i is the inverse of the selection probability, perhaps adjusted for non-response and post-stratification. Setting $\hat{S}(\beta)$ equal to $\mathbf{0}$ gives us our estimator, $\hat{\beta}$. We could use linearization or the jackknife directly on $\hat{\beta}$ to get standard errors. Alternatively, we can expand $\hat{S}(\hat{\beta})$ about the true value, β , and obtain as our estimated covariance matrix the “sandwich” estimator

$$\hat{\text{Cov}}\{\hat{\beta}\} \approx \mathbf{J}(\hat{\beta})^{-1} \hat{\text{Cov}}\{\hat{S}(\hat{\beta})\} \mathbf{J}(\hat{\beta})^{-1}, \quad (4)$$

where $\mathbf{J}(\beta) = -\partial \hat{S} / \partial \beta^T = \sum_{\text{sample}} w_i p_i(\mathbf{x}_i; \beta) p_0(\mathbf{x}_i; \beta) \mathbf{x}_i \mathbf{x}_i^T$ with $p_0 = 1 - p_1$. Since $\hat{S}(\beta)$ is a vector of totals, $\hat{\text{Cov}}\{\hat{S}(\beta)\}$ should be available as a matter of course for any standard design. Most major statistical packages (for

example, SAS (PROC SURVEYLOGISTIC), SPSS (CSLOGISTIC), STATA (SVY:LOGIT), SUDAAN (LOGISTIC)) can handle logistic regression with complex sampling and weighting routinely these days. Thus producing weighted estimates and making associated inferences is reasonably straightforward.

Strictly speaking, the selection probabilities will themselves often be random variables in our model-based framework, based on a finite population that we assume is generated from the model. We can account for this by using the results in Rao (1973), but the correction is of order $1/N$ and can be ignored in most large studies.

The downside of weighting in general is that it tends to be inefficient when the weights are highly variable. (A rule-of-thumb sometimes suggested is that w_{\max} / w_{\min} should be no more than 10.) In case-control studies, the variation in weights is about as extreme as it can get. For instance, the ratio of w_{\max} to w_{\min} is approximately 300:1 in Example 1 and approximately 1,000:1 in Example 2. Even more extreme ratios are not uncommon. No experienced survey sampler would be surprised to find that weighting is not very efficient under these circumstances.

Can we do something more efficient? The answer is certainly “Yes” in some special cases. Fully efficient likelihood methods have been developed in situations where both cases and controls are drawn using simple or stratified random sampling and these can be very much more efficient than weighted methods. We review these results in the next section.

5. Review: Simple Case

We start with the very simplest case where cases and controls are selected by simple random sampling and we have no population information about any of the covariates at the design stage. Here fully efficient semi-parametric maximum-likelihood procedures are well-developed. Moreover, these methods are very simple to implement using standard software (Prentice and Pyke 1979). (The methods are *semi-parametric* because the full likelihood depends on the unknown distribution of the covariates and we do not want to model this in general.)

It turns out that all we have to do is fit model (1) using a standard logistic regression program without any weighting at all. More specifically, solving the unweighted equation

$$\sum_{\text{sample}} \mathbf{x}_i (y_i - p_i(\mathbf{x}_i; \beta)) = 0, \quad (5)$$

produces efficient estimates of all the coefficients except the intercept. Perhaps more importantly, all the standard errors and resulting inferences that we get from the standard program are also valid, again with the exception of anything

involving the intercept. It is simple enough to correct inferences involving the intercept provided that we know the ratio of the two sampling fractions but we are often only interested in the other coefficients anyway.

The results extend directly to stratified random sampling, provided that separate intercepts for each stratum are included in the model. Again efficient semi-parametric estimators of all coefficients except the stratum intercepts can be obtained simply by running the data through an ordinary (unweighted) logistic regression program. Again, the estimated standard errors and associated inferences are also valid. As with simple random sampling, we can correct the results for the stratum intercepts provided that we know the stratum sampling fractions but, again, these are usually of minor interest.

Thus in these simple situations, maximum likelihood estimates are simpler to compute than the weighted estimates, as well as being more efficient. How much more efficient are they? This depends on the number of covariates, the magnitude of their coefficients and the ratio of the sampling fractions, but the difference is often substantial. (The weighted estimates are about 50% efficient in Example 2 of the introduction, for example, and less than 20% efficient in the brain cancer example we look at in Section 8. Lawless, Kalbfleisch and Wild 1999 discuss situations where the efficiency is even lower than this.)

Finally, we note that the maximum likelihood estimates have yet another advantage over weighted estimates: they tend to have much better small sample performance, especially in situations where the efficiency of the weighted estimates is low. Essentially, weighting results in a reduction in the effective sample size and it is this effective sample size that governs when the asymptotic theory starts to give a good approximation. (See Scott and Wild 2001a for more details.) Clearly we can pay a very heavy price for a rigid adherence to population weights.

6. More Complex Sampling

In both the examples in Section 2, the controls were obtained from a complex multi-stage survey rather than a simple random sample. As we noted in the introduction, this is increasingly common in large scale case-control studies. (Occasionally, as in Example 1, the cases are also selected using a complex sampling scheme.) It is possible to derive semi-parametric efficient estimators for stratified multistage sampling, assuming that primary sampling units are selected independently within strata (which is the assumption that all the computer packages are making with the survey-weighted approach anyway), but this requires us to build multivariate models for the vector of responses within a primary sampling unit. Details can be found in Neuhaus,

Scott and Wild (2002, 2006). Unless we are interested in the within-cluster structure in its own right (as in the family case-control studies considered in Section 9, for example), this requires far too much effort for it to be practicable, certainly for routine analysis.

Can we do something simpler without losing too much efficiency? Weighted estimates are always available, of course. However, they are just as inefficient with complex designs as they are in the simple case considered in the previous section. It turns out that we can do considerably better without too much extra complication.

Return for a moment to the situation of the previous section where we have a simple random sample of size n_1 from the case stratum and an independent simple random sample of size n_0 from the control stratum. Here all units in Stratum ℓ have weight $w_i \propto W_\ell / n_\ell$, where W_ℓ denotes the proportion of the population in the stratum, for $\ell = 0, 1$. If we divide throughout by N and set $p_0(\mathbf{x}; \beta) = 1 - p_1(\mathbf{x}; \beta)$, then we can re-write equation (3) for the weighted estimator in the form

$$\frac{\sum_{\text{cases}} \mathbf{x}_i p_0(\mathbf{x}_i; \beta)}{n_1} - \frac{\sum_{\text{controls}} \mathbf{x}_i p_1(\mathbf{x}_i; \beta)}{n_0} = \mathbf{0}. \quad (6)$$

Similarly, we can write equation (5) for the efficient maximum likelihood estimator in the form

$$\omega_1 \frac{\sum_{\text{cases}} \mathbf{x}_i p_0(\mathbf{x}_i; \beta)}{n_1} - \omega_0 \frac{\sum_{\text{controls}} \mathbf{x}_i p_1(\mathbf{x}_i; \beta)}{n_0} = \mathbf{0}, \quad (7)$$

where $\omega_\ell = n_\ell / (n_0 + n_1)$, for $\ell = 0, 1$. Both these are special cases of the general set of estimating equations

$$\lambda_1 \frac{\sum_{\text{cases}} \mathbf{x}_i p_0(\mathbf{x}_i; \beta)}{n_1} - \lambda_0 \frac{\sum_{\text{controls}} \mathbf{x}_i p_1(\mathbf{x}_i; \beta)}{n_0} = \mathbf{0}. \quad (8)$$

As $n_0, n_1 \rightarrow \infty$, under mild conditions on the way that the finite population is generated from the superpopulation the solution of (8) converges almost surely to the solution β^* of

$$\lambda_1 E_1 \{ \mathbf{X} p_0(\mathbf{X}; \beta^*) \} - \lambda_0 E_0 \{ \mathbf{X} p_1(\mathbf{X}; \beta^*) \} = \mathbf{0}, \quad (9)$$

where $E_\ell \{ \cdot \}$ denotes the conditional expectation given that $Y = \ell$ for $\ell = 0, 1$. If model (1) is true, then equation (8) has solution $\beta_1^* = \beta_1$ and $\beta_0^* = \beta_0 + b_\lambda$ with $b_\lambda = \log(\lambda_1 W_0 / \lambda_0 W_1)$ for any positive λ_0, λ_1 (see Scott and Wild 1986 for details of the proof). Thus the solution to equation (8) produces consistent estimators of all the regression coefficients apart from the constant term for any $\lambda_\ell > 0$ ($\ell = 0, 1$). As in the simple case, it is easy to correct the inferences about the constant term, provided that we know the proportion of cases in the population.

Now turn to more complex sampling schemes. Since the left hand side of equation (9) just involves two subpopulation means, we can still estimate these means for any standard survey design. This suggests an estimator, $\hat{\beta}_\lambda$ say, for general sampling schemes satisfying

$$\hat{S}_\lambda(\beta) = \lambda_1 \hat{\mu}_1(\beta) - \lambda_0 \hat{\mu}_0(\beta) = 0, \quad (10)$$

where $\hat{\mu}_\ell(\beta)$ is the sample estimator of the subpopulation mean $E_\ell\{\mathbf{X}(1 - p_\ell(\mathbf{X}; \beta))\}$ ($\ell = 0, 1$). The covariance matrix of $\hat{\beta}_\lambda$ can then be obtained by standard linearization arguments. This leads to an estimated ('sandwich') covariance matrix

$$\hat{\text{Cov}}\{\hat{\beta}_\lambda\} \approx \mathbf{J}_\lambda(\hat{\beta}_\lambda)^{-1} \hat{\text{Cov}}\{\hat{S}_\lambda(\hat{\beta}_\lambda)\} \mathbf{J}_\lambda(\hat{\beta}_\lambda)^{-1}, \quad (11)$$

with $\mathbf{J}_\lambda(\beta) = (-\partial \hat{S}_\lambda(\beta) / \partial \beta^T)$ and $\hat{\text{Cov}}\{\hat{S}_\lambda(\beta)\} = \lambda_1^2 \hat{\text{Cov}}\{\hat{\mu}_1(\beta)\} + \lambda_0^2 \hat{\text{Cov}}\{\hat{\mu}_0(\beta)\}$. Here, $\hat{\text{Cov}}\{\hat{\mu}_\ell(\beta)\}$ denotes the usual survey estimate which should be available routinely for any standard survey design since $\hat{\mu}_\ell(\beta)$ is just an estimated mean.

All of this can also be carried out straightforwardly in any package that can handle logistic regression for complex survey designs simply by specifying the appropriate vector of weights. More specifically, suppose that

$$\hat{\mu}_\ell(\beta) = \frac{\sum_{i \in S_\ell} w_i \mathbf{x}_i (y_i - p_1(\mathbf{x}_i; \beta))}{\sum_{i \in S_\ell} w_i}, \quad (12)$$

where S_1 denotes the case subpopulation (*i.e.*, the set of all units with $Y = 1$) and S_0 denotes the control subpopulation (the set of all units with $Y = 0$). Then the estimating equation (9) can be written in the form

$$\hat{S}_\lambda(\beta) = \sum_{\text{sample}} w_i^* \mathbf{x}_i (y_i - p_1(\mathbf{x}_i; \beta)) = 0, \quad (13)$$

with $w_i^* \propto \lambda_\ell w_i / \sum_{i \in S_\ell} w_i$ for units in S_ℓ ($\ell = 0, 1$). In other words, we simply have to scale the case weights and control weights separately so that the sum of the case weights is proportional to λ_1 and the sum of the control weights is proportional to λ_0 and put them, along with the usual specification of the design structure (strata, primary sampling units), into our program of choice. Note that the choice of proportionality constant does not affect the result.

We still have to decide on good values for λ_1 and λ_0 . We can often get substantial gains using sample weights ($\lambda_\ell = n_\ell / n$) compared with using population weights ($\lambda_\ell = W_\ell$). Scott and Wild (2002) report efficiency gains of 50% or more in Example 2 and in simulations based on that population. The gains became larger as the strength of the relationship increased, and as the effect of clustering increased. Moreover the coverage of confidence intervals

was closer to the nominal value for sample weighting in the simulations.

Using sample weights is the most efficient possible strategy when we have simple random samples of cases and controls but for more complex schemes using the sample weights will no longer be fully efficient. We might expect weights based on some form of equivalent sample sizes to perform better. This does indeed produce some gain in efficiency in some limited simulations reported in Scott and Wild (2001a). However, the gains are relatively small, at least when the control sample design effect is less than 2, since $\text{Cov}\{\hat{\beta}_\lambda\}$ is very flat as a function of λ near its minimum. Considerations of robustness that we discuss in Section 8 are possibly more important in the choice of λ .

The gains from sample weighting may depend very much on the particular problem under examination. Korn and Graubard (1999, page 327) comment that, in their experience, the sample weighting strategy rarely produces big gains in efficiency. Obviously more work, both empirical and theoretical, is needed here. In any event, it would seem prudent to fit the model using both sample weights and population weights routinely. If the coefficient estimates are similar, then we can make a judgement based on the estimated standard errors. However, significant differences in the coefficient estimates indicate that the model has been mis-specified. If we are unable to fix up the deficiencies in the model, then we need to think very carefully about just what it is that we are trying to estimate. We look at this again in Section 8.

7. Stratified Sampling

The compromise suggested in the previous section (*i.e.*, use standard survey weighting within the subpopulations defined by case/control status but combine the subpopulations using sample proportions) seems to work reasonably well in practice but it is all completely *ad hoc*. Could we do better with a more systematic approach?

In the special case of stratified random sampling, where independent case-control samples are taken within each stratum, fully efficient procedures are well-developed and easy to implement. In particular, if our model includes a separate intercept for each stratum, then ordinary unweighted logistic regression (with a simple adjustment for the stratum intercepts if they are wanted) is the efficient semi-parametric maximum likelihood procedure (Prentice and Pyke 1979). It is reasonably straightforward to extend this to more general stratified designs. Our model is now

$$\text{logit}\{P(Y = 1 \mid \mathbf{x}, \text{Stratum } h)\} = \beta_{0h} + \mathbf{x}^T \beta_1, \quad (14)$$

and the stratified equivalent of the estimating equation (7) is

$$\sum_h \left(\lambda_{1h} \frac{\sum \mathbf{x}_i p_{0h}(\mathbf{x}_i; \beta)}{n_{1h}} - \lambda_{0h} \frac{\sum \mathbf{x}_i p_{1h}(\mathbf{x}_i; \beta)}{n_{0h}} \right) = \mathbf{0}. \quad (15)$$

As $n_{0h}, n_{1h} \rightarrow \infty$, the solution of (7) converges almost surely to the solution of

$$\sum_h (\lambda_{1h} E_{1h} \{\mathbf{X} p_{0h}(\mathbf{X}; \beta)\} - \lambda_{0h} E_0 \{\mathbf{X} p_{1h}(\mathbf{X}; \beta)\}) = \mathbf{0}, \quad (16)$$

with the obvious extension of the notation from the unstratified case. If model (13) is true, then equation (8) has solution $\beta_1^* = \beta_1$ and $\beta_{0h}^* = \beta_{0h} + b_{\lambda h}$ with $b_{\lambda h} = \log(\lambda_{1h} W_{0h} / \lambda_{0h} W_{1h})$. Since equation (14) only involves stratum means, we can estimate them easily using the data coming from any reasonable survey design, for example by

$$\hat{\mu}_{th}(\beta) = \frac{\sum_{i \in S_{th}} w_{ih} \mathbf{x}_{ih} (y_{ih} - p_1(\mathbf{x}_{ih}; \beta))}{\sum_{i \in S_{th}} w_{ih}}.$$

Substituting these estimators in place of the sample means in equation (14) leads to the estimating equation

$$\hat{S}_\lambda(\beta) = \sum_h \sum_{i \in S_{th}} w_{ih}^* \mathbf{x}_i (y_i - p_{1h}(\mathbf{x}_i; \beta)) = \mathbf{0}, \quad (17)$$

with $w_{ih}^* \propto \lambda_{th} w_{ih} / \sum_{i \in S_{th}} w_{ih}$ for units in S_{th} ($\ell = 0, 1$; $h = 1, \dots, H$). This can be fitted in any standard survey program by including these weights and the appropriate design information. Note that we need to be careful about how we include the so-called ‘strata’ in the design specification. If primary sampling units are nested within the ‘strata’, as with the geographical locations in Example 1, there is no problem and the strata should be included in the standard way. However, if the primary sampling units cut across the ‘strata’, as with age in Example 1 and age and ethnicity in Example 2, then these are not strata in the usual survey sampling sense. They should not be included in the design specifications but simply handled through the weights.

Sometimes we want to model the contribution of the stratum variables using some smooth parametric curve rather than including them through dummy variables. For example, we might well want to include a linear function of age in our model in both Examples 1 and 2. The survey weighted method and the compromise weighting suggested in Section 6 both apply directly and no new theory is needed. More efficient methods are not nearly so simple, however. Fully efficient methods have been developed in the case where simple random samples of cases and controls are drawn within each of the strata (see Scott and Wild

1997, and Breslow and Holubkov 1997) but the resulting estimating equations are not linear combinations of stratum means and there is no obvious way of generalizing them to more complex sampling schemes. There is a slightly less efficient way that does extend easily, however. If we modify model (14) by including $b_{\lambda h} = \log(\lambda_{1h} W_{0h} / \lambda_{0h} W_{1h})$ as an offset, i.e., we set

$$\text{logit}\{P^*(Y = 1 \mid \mathbf{x}, \text{Stratum } h)\} = b_{\lambda h} + \beta_{0h} + \mathbf{x}^T \beta_1, \quad (18)$$

then equation (15) produces consistent, fully efficient, estimates of all the coefficients including β_{0h} ($h = 1, \dots, H$). Including the same offsets in models where there is no β_{0h} term and the \mathbf{x} vector includes functions of the stratifying variable produces consistent estimators of all the coefficients with typically high (although not full) efficiency (see Fears and Brown 1986, and Breslow and Cain 1988). This generalizes to arbitrary designs immediately. We just use equation (16) with p_{1h} replaced by p_{1h}^* defined by setting $\text{logit}(p_{1h}^*) = b_{\lambda h} + \mathbf{x}^T \beta$. Then any survey program that caters for offsets can be used to fit the model and provide estimated standard error, etc.

How much extra efficiency do we get in this case? We have carried out a number of simulations, some of which are reported in Scott and Wild (2002). Most of the scenarios are based on the meningitis study in Example 2 and we set the ratio of the largest to smallest stratum sampling fraction in the control sample at about 10:1. Without any clustering, the gain in efficiency from using the offset method (which is full maximum likelihood in this case) compared to the *ad hoc* procedure was never more than 10%. The relative efficiencies stayed about the same as clustering that cut across strata was introduced. When clustering nested within strata was introduced, the gains disappeared progressively as the design effect increased and the *ad hoc* procedure actually became more efficient than the offset method when the design effect reached about 1.5.

As we stated earlier, it is possible to produce fully efficient semi-parametric estimators if we are willing to model the dependence structure within primary sampling units. We have begun to carry out some simulation. The early results suggest that the extra work involved in the modeling will almost never be worth the effort if we are only interested in the parameters of the marginal model (1). Our tentative conclusion is that, the *ad hoc* partially weighted procedures (with sample weights) are simple to use and work well enough for most practical purposes in the range covered by our experience but this is another area where more empirical work is needed yet. We note, however, that there are some problems, like the case-control family design discussed in Section 9, where the within-cluster behavior is of interest in its own right. These require more sophisticated methods.

8. Robustness

There must be a catch somewhere. What if the model is not correct? What price do we pay for efficiency then?

By its construction, the population-weighted estimator is always estimating the linear logistic approximation that we would get if we had data from the whole population. By contrast, what the more efficient sample-weighted estimator is estimating depends on the particular sample sizes used. Some people would regard this alone as a strong enough reason for using the population weighted estimator and I suspect that very few people would regard it as completely satisfactory to have the target of their inference depend on the arbitrary choice of sample size.

Our general estimator $\hat{\beta}_\gamma$ satisfying (10) converges to the solution of equation (9), \mathbf{B}_γ say, with $\gamma = \lambda_0/(\lambda_0 + \lambda_1)$, which depends on the true model and distribution of the covariates, as well as on γ . In Scott and Wild (2002), we looked at what happens to \mathbf{B}_γ under mild deviations from the assumed model. (We are interested in small deviations since large ones should be picked up by routine model-checking procedures and the model then improved.) For simplicity, suppose that we fit a linear model with a single explanatory variable for the log odds ratio but that the true model is quadratic, say

$$\text{logit}\{P(Y = 1 | x)\} = \beta_0 + \beta_1 x + \delta x^2 \quad (19)$$

with δ small.

Obviously, the actual slope on the logit scale, $\beta_1 + 2\delta x$, changes as we move along the curve. For any $0 < \gamma < 1$, $\mathbf{B}_{\gamma 1}$ is equal to the actual slope at some point along the curve. Denote this value by $x = x_\gamma$. Let x_0 be the expected value of x in the control population and let x_1 the expected value of x in the case population. We shall assume that $\beta_1 > 0$ so that $x_0 < x_1$. It turns out that x_γ always lies between x_0 and x_1 and that x_γ increases as γ increases from 0 to 1. Recall that survey weighting corresponds to $\gamma = W_0$ and sample weighting to $\gamma = \omega_0 = n_0/n$. Typically, W_0 is much larger than ω_0 so that survey weighting gives an estimate of the slope at larger values of x , where the probability of a case is higher, while the slope estimated from sample weighting is closer to the average value of x in the population. Figure 1, adapted from Scott and Wild (2002), illustrates the position in two scenarios, one with positive curvature and one with negative, based roughly on Example 2. The value of δ is chosen so that it would be detected with a standard likelihood ratio test about 50% of the time if we took simple random samples of $n_0 = n_1 = 200$ from the population.

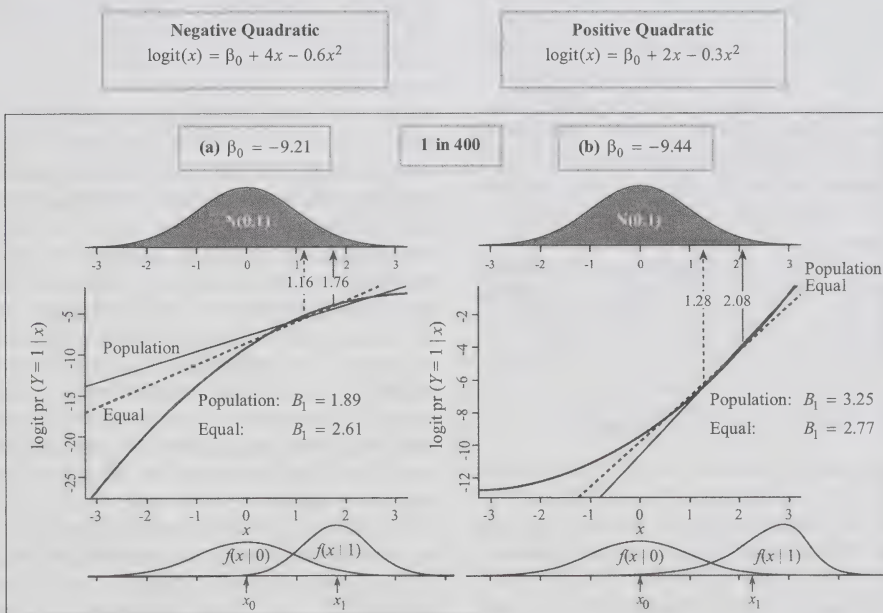


Figure 1. Comparison of population and equal weights.

In both scenarios, the value of β_0 is set so that the proportion of cases in the population is 1 in 400, *i.e.*, so that $W_0 = 0.9975$. The overall density of x is shown at the top of the graph and the conditional densities for cases and controls are shown at the bottom. Values of x_γ and $\mathbf{B}_{\gamma 1}$ are shown for $\gamma = W_0$ (labeled “population”) and $\gamma = 0.5$ (labeled “equal”). The latter value corresponds to sample weighting if we draw equal numbers of cases and controls. Clearly, survey weighting is estimating the appropriate slope for values of x further out in the upper tail of the distribution (*i.e.*, for individuals at higher risk) than equal weighting in both scenarios.

Note that if we took simple random samples of $n_0 = n_1 = 200$ from the population in Figure 1 (a), the relative efficiency of survey weighting is only about 16%, and the small sample bias is 0.24. In this case, even if we take the population value as our target, the survey weighting leads to a larger mean squared error than sample weighting.

More results are given in Scott and Wild (2002) where we also look at the effect of omitted covariates. This turns out to have a similar, but somewhat smaller, effect to omitting a quadratic term.

Which is the right value of γ to use? That clearly depends on what we want to use the resulting model for. If our primary interest is in using the model for estimating odds ratios at values of x where the probability of a case is higher, and the sample is large enough so that variance and small sample bias are less important, we might use population weights. For smaller sample sizes, or if we are interested in values of x closer to the population mean, sample weights would be better. A value intermediate between population weighting and sample weighting might sometimes be a sensible compromise. For example trimming the weights to 10:1 (*i.e.*, setting $\gamma \approx 0.91$) in the example, instead of 1:1 (sample weighting) or 400:1 (population weighting), leads to an efficiency of 70% and a small sample bias of 0.04. The corresponding values for population weighting were 16% and 0.24. The value of $x_{0.91}$ lies almost exactly half way between $x_{0.5}$ and $x_{0.9975}$.

9. Case-Control Family Studies

If we are primarily interested in the parameters of the marginal model (1), then the methods that we have discussed in previous sections are simple to implement and reasonably efficient. Fully efficient methods require building parametric models for the within-cluster dependence and the extra effort that this would entail is rarely worthwhile. However, there are situations where the dependence structure is of interest in its own right. In particular, it has become increasingly common for genetic epidemiologists to augment data from a standard case-control study with response and covariate

information from family members, in an attempt to gain information on the role of genetics and environment. This can be regarded as a stratified cluster sample, with families as clusters, and the intra-cluster structure is of the primary focus of attention here. The following example is fairly typical.

Example 3.

Wrensch, Lee, Miike, Newman, Barger, Davis, Wiencke and Neuhaus (1997) conducted a population-based case-control study of glioma, the most common type of malignant brain tumor, in the San Francisco Bay Area. They collected information on all cases of glioma that were diagnosed in a specified time interval and on a comparable sample of controls obtained through random digit dialing. They also collected brain tumor status and covariate information from family members of the participants in the original case-control sample. There were 476 brain cancer case families and 462 control families in the study.

We could use the methods that we have been discussing to fit a marginal model for the probability of becoming a glioma victim but a major interest of the researchers was the estimation of within-family characteristics. One way of approaching this would be to fit a mixed logistic model with one or more random family effects.

Note that, strictly speaking, the original sampling scheme in Example 3 is not included in this case-control set-up. The stratification here is related to the response variable but not completely determined by it. Stratum 1 contains the 476 families with a case diagnosed in a particular small time interval while Stratum 2 contains the remaining 1,942,490 families, some of which contain brain cancer victims.

In Neuhaus *et al.* (2006) we develop efficient semi-parametric methods for stratified multi-stage sampling in situations where the stratification depends on the response, possibly in an unspecified way that has to be modeled, and observations within a primary sampling unit are related through some parametric model. The estimates require the solution of $p + 1$ estimating equations, where p is the dimension of the parameter vector. The covariance matrix can also be estimated in a straightforward way using an analogue of the inverse observed information matrix. The whole procedure can be implemented using any reasonably general maximization routine but this still requires some computing expertise.

We could also fit the same models using survey weighted estimators, which has the big advantage of requiring no specialist software. In our example, case families would have weight 1 and control families would have weight $1,942,490/462 \approx 4,200$. With such a huge disparity, we might expect the weighted estimates to be very inefficient indeed. Unfortunately it turned out to be almost impossible to fit an interesting model for which the weighted estimates

converged. One problem is that the weighted estimates are based almost entirely on the control sample and there is very little information about family effects in the control families. (Another problem is that we did not have information on age for family members and any model without age was grossly mis-specified!) For this reason, we had to resort to simulation which is far from complete at this stage. It seems, however, that the efficiency of weighted estimates is less than 10% of the efficient semiparametric estimates here. More details are given in Neuhaus *et al.* (2002, 2006).

Although our simulations are at a very early stage, it is possible to draw a few tentative conclusions. The main one is that within-family quantities are very poorly estimated, even using fully-efficient procedures. Case-control family designs, where the information on family members is obtained as an add-on to a standard case-control design, simply do not contain enough information to estimate the parameters of interest to genetic epidemiologists unless the associations are extremely (even unrealistically) strong. (I should note that not all genetic epidemiologists would agree with this.) More efficient variants are possible, however. For example, if we can identify families containing more than one case, then it is possible to get much greater efficiency by heavily over-sampling such families. In essence, we would be taking the family as the sampling unit, defining a 'case family' as one containing multiple individual cases and then taking a case-control sample of families. This is an important area where a lot of work still needs to be done.

10. Conclusion

The population-based case-control study is one of those subjects where practice has forged ahead of theory. As far as I know, the only book that discusses the topic in any depth is Korn and Graubard (1999, Chapter 9). One aspect that has received a reasonable amount of theoretical attention in the literature is stratification. Efficient procedures for incorporating stratifying variables in the analysis have been developed by Scott and Wild (1997), Breslow and Holubkov (1997), and Lawless *et al.* (1999), among others, when the variables can take only a finite set of values. Breslow and Chatterjee (1999) have considered how best to use such information at the design stage. The extension of all this (both analysis and design) to situations where we have information on continuous variables such as age for all members of the population is an area that still needs work. Much less has been written on the effect of clustering, even though multi-stage sampling is in common use. Exceptions are Graubard *et al.* (1989), Fears and Gail (2000) and Scott and Wild (2001a). Perhaps this paper might stimulate more work on an important topic. In particular, since the essence of the problem boils down to estimating two population

means (see equation (8)), it should be possible to transfer a lot of the expertise about efficient survey design across to this problem.

Acknowledgements

I would like to thank the referees and Barry Graubard and Graham Kalton, whose thoughtful discussion of an early version of this paper helped my understanding of the subject considerably. Finally, I would give special thanks to my long term collaborators Chris Wild, with whom almost all the work underlying this paper was done, and Jon Rao, with whom I learnt essentially everything that I know about the analysis of survey data.

References

- Baker, M., McNicholas, A., Garrett, N., Jones, N., Stewart, J., Koberstein, V. and Lennon, D. (2000). Household crowding: A major risk factor for epidemic meningococcal disease in Auckland children. *Pediatric Infectious Disease Journal*, 19, 983-990
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Breslow, N.E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, 91, 14-28.
- Breslow, N.E. (2004). Case-control studies. In *Handbook of Epidemiology*. (Eds. W. Aherns and I. Pigeot). New York: Springer. 287-319.
- Breslow, N.E., and Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75, 11-20.
- Breslow, N.E., and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Applied Statistics*, 48, 457-468.
- Breslow, N.E., and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase outcome-dependent sampling. *Journal of the Royal Statistical Society, B*, 59, 447-461.
- Brogan, D.J., Denniston, M.M., Liff, J.M., Flagg, E.W., Coates, R.J. and Brinton, L.A. (2001). Comparison of telephone sampling and area sampling: Response rates and within-household coverage. *American Journal of Epidemiology*, 153, 1119-1127.
- Cosslett, S.R. (1981). Maximum likelihood estimation for choice-based samples. *Econometrica*, 49, 1289-1316.
- DiGaetano, R., and Waksberg, J. (2002). Trade-offs in the development of a sample design for case-control studies. *American Journal of Epidemiology*, 155, 771-775.
- Fears, T.R., and Brown, C.C. (1986). Logistic regression models for retrospective case-control studies using complex sampling procedures. *Biometrics*, 42, 955-960.
- Fears, T.R., and Gail, M.H. (2000). Analysis of a two-stage case-control study with cluster sampling of controls: Application to non-melanoma skin cancer. *Biometrics*, 56, 190-198.

- Graubard, B.I., Fears, T.R. and Gail, M.H. (1989). Effects of cluster sampling on epidemiologic analysis in population-based case-control sampling. *Biometrics*, 45, 1053-1071.
- Hartge, P., Brinton, L.A., Rosenthal, J.F., Cahill, J.I., Hoover, R.N. and Waksberg, J. (1984). Random digit dialing in selecting a population-based control group. *American Journal of Epidemiology*, 120, 825-833.
- Hartge, P., Brinton, L.A., Cahill, J.I., West, D., Hauk, M., Austin, D., Silverman, D. and Hoover, R.N. (1984). Design and methods in a multi-center case-control interview study. *American Journal of Public Health*, 74, 52-56.
- Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.
- Lawless, J.F., Kalbfleisch, J.D. and Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society*, B, 61, 413-38.
- Lee, A.J., Scott, A.J. and Wild, C.J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika*, 95 (to appear).
- Manski, C.F., and McFadden, D. (Eds) (1981). *Structural Analysis of Discrete Data with Econometric Applications*. New York: John Wiley & Sons, Inc.
- Miettinen, O.S. (1985). The case-control study: Valid selection of subjects. *American Journal of Epidemiology*, 135, 1042-1050.
- Prentice, R.L., and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.
- Neuhaus, J., Scott, A.J. and Wild, C.J. (2002). The analysis of retrospective family studies. *Biometrika*, 89, 23-37.
- Neuhaus, J., Scott, A.J. and Wild, C.J. (2006). Family-specific approaches to the analysis of retrospective family data. *Biometrics*, 62, in press.
- Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.
- Rao, J.N.K., Scott, A.J. and Skinner, C.J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, 8, 1059-1070.
- Scott, A.J., and Wild, C.J. (1986). Fitting logistic models under case-control or choice-based sampling. *Journal of the Royal Statistical Society*, B, 48, 170-182.
- Scott, A.J., and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 83, 57-72.
- Scott, A.J., and Wild, C.J. (2001a). The analysis of clustered case-control studies. *Applied Statistics*, 50, 57-71.
- Scott, A.J., and Wild, C.J. (2001b). Fitting regression models to case-control data by maximum likelihood. *Journal of Statistical Planning and Inference*, 96, 3-27.
- Scott, A.J., and Wild, C.J. (2002). On the robustness of weighted methods for fitting model to case-control data by maximum likelihood. *Journal of the Royal Statistical Society*, B, 64, 207-220.
- Wacholder, S., McLaughlin, J.K., Silverman, D.T. and Mandel, J.S. (1991). Selection of controls in case-control studies. I. Principles. *American Journal of Epidemiology*, 135, 1019-1028.
- Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.
- Waksberg, J. (1998). Random digit dialing sampling for case-control studies. In *Encyclopedia of Biostatistics*. (Eds. P.Armitage and T. Colton). New York: John Wiley & Sons, Inc., 3678-3682.
- Wrensch, M., Lee, M., Miike, R., Newman, B., Barger, G., Davis, R., Wiencke, J. and Neuhaus, J. (1997). Familial and personal medical history of cancer and nervous system conditions among adults with glioma and controls. *American Journal of Epidemiology*, 145, 581-93.

Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors

Phillip S. Kott¹

Abstract

Calibration weighting can be used to adjust for unit nonresponse and/or coverage errors under appropriate quasi-randomization models. Alternative calibration adjustments that are asymptotically identical in a purely sampling context can diverge when used in this manner. Introducing instrumental variables into calibration weighting makes it possible for nonresponse (say) to be a function of a set of characteristics other than those in the calibration vector. When the calibration adjustment has a nonlinear form, a variant of the jackknife can remove the need for iteration in variance estimation.

Key Words: Prediction model; Quasi-randomization model; Quasi-randomization consistent; Instrumental variable; Generalized raking.

1. Introduction

Calibration weighting was originally developed as a method for reducing sampling errors while retaining randomization consistency. Deville and Särndal (1992) demonstrated that many alternative forms of calibration weighting are asymptotically identical in the sampling context. This lead to a breakthrough in our understanding of common weight adjustment methods like raking that do not appear in generalized-regression (GREG) estimator format.

Folsom and Singh (2000) showed that calibration weighting can also be used to adjust for known coverage errors and/or unit nonresponse under appropriate quasi-randomization models. Their work is not in the refereed literature. The heart of this article repeats key results in Folsom and Singh including a necessary modification of the Deville-Särndal approach to model variance/randomization mean-squared-error estimation in this expanded context. An earlier, strictly linear version of calibration weighting for unit-nonresponse adjustment can be found in Fuller, Loughin and Baker (1994). See also Lundström and Särndal (1999).

A distinction is drawn between the prediction model usually underpinning calibration and the quasi-randomization model in Folsom and Singh. Unlike in Folsom and Singh, however, both properties are explored here. Furthermore, the explanatory variables in the quasi-randomization model are allowed to differ from the calibration variables. This is likewise allowed in Lundström and Särndal.

A new jackknife is proposed which is analogous to the Deville-Särndal linearization variance estimator. It employs replicate weights computed in one step even though the calibration weights themselves may be determined iteratively.

After introducing the popular notion of calibration weighting, Section 2 provides a review of the GREG special

case in a purely sampling context. Section 3 describes Estevao and Särndal's (2000) extension of calibration weighting in its linear form to include instrumental variables. Section 4 expands Deville and Särndal's treatment of calibration weighting to include the possibility of instrumental variables. Section 5 reviews variance/mean squared error estimation, proposing a new jackknife for certain designs. Section 6 describes how calibration weighting can be used to adjust for nonresponse. In this context, alternative functional forms of calibration weighting need no longer be asymptotically identical. Section 7 discusses quasi-randomization models for coverage errors, that is, frame under- or over-coverage. Section 8 contains a small empirical example supporting the new jackknife. Section 9 provides a discussion of alternative approaches and areas for future research.

2. Calibration Weighting and the GREG Estimator

Suppose we knew the selection probability, π_k , for each sample element k in the sample S . We can estimate any population total, $T_y = \sum_U y_k$, where U denotes the population, with the expansion estimator $t_{y_E} = \sum_S y_k / \pi_k = \sum_U y_k I_k / \pi_k$, where $I_k = 1$ when $k \in S$ and 0 otherwise. Treating the I_k as random variables, it is easy to see that t_{y_E} is an unbiased estimator for T_y . Properties arising when the I_k are treated as random variables are called *randomization-based*. We can also write $t_{y_E} = \sum_U a_k y_k = \sum_S a_k y_k$, where $a_k = I_k / \pi_k$ is the *sampling weight* of element k .

Deville and Särndal (1992) coined the term "calibration estimator" to describe an estimator of the form $t_{y_CAL} = \sum_S w_k y_k$, where $\sum_S w_k \mathbf{x}_k = \sum_U \mathbf{x}_k = T_{\mathbf{x}}$ for some

1. Phillip S. Kott, National Agricultural Statistics Service, 3251 Old Lee Highway, Fairfax, VA 22030, U.S.A. E-mail: pkott@nass.usda.gov.

row vector of auxiliary variables, $\mathbf{x}_k = (x_{1k}, \dots, x_{pk})$, about which $T_{\mathbf{x}}$ is known. Since there is generally a continuum of sets $\{w_k | k \in S\}$ that satisfy the calibration equation:

$$\sum_{k \in S} w_k \mathbf{x}_k = T_{\mathbf{x}}, \quad (1)$$

Deville and Särndal required that the difference between the set of weights, $\{w_k | k \in S\}$, satisfying equation (1) and $\{a_k | k \in S\}$ minimize some loss function.

An alternative approach to survey sampling treats the y_k as random variables satisfying the linear prediction model:

$$y_k = \mathbf{x}_k \boldsymbol{\beta} + \varepsilon_k, \quad (2)$$

where $E(\varepsilon_k | \{\mathbf{x}_g, I_g | g \in U\}) = 0$ for all $k \in U$. By conditioning this expectation on the I_g , we are assuming the sampling mechanism can be ignored. This is a crucial, and sometimes unreasonable, aspect of the (prediction) model-based framework.

It is easy to see that t_{y_CAL} is an unbiased estimator for T_y under the model in the sense that $E_{\varepsilon}(t_{y_CAL} - T_y) = 0$ (suppressing the conditioning for notational convenience); the subscript ε refers to treating the ε_k as random variables (and the I_k as fixed constants).

For our purposes, the general(ized) regression or GREG estimator has the form:

$$t_{y_GREG} = t_{y_E} + \left(T_{\mathbf{x}} - \sum_{k \in S} a_k \mathbf{x}_k \right) \left(\sum_{k \in S} c_k a_k \mathbf{x}'_k \mathbf{x}_k \right)^{-1} \sum_{k \in S} c_k a_k \mathbf{x}'_k y_k, \quad (3)$$

where c_k is an arbitrary constant which may or may not be a function of \mathbf{x}_k , and $\lim_{N \rightarrow \infty} \sum_{k \in U} c_k \mathbf{x}'_k \mathbf{x}_k / N = \Lambda$ is a positive definite matrix, where N is the size of U . This last condition means that $\sum_S c_k a_k \mathbf{x}'_k \mathbf{x}_k$ will usually be invertible in practice. We will assume that it is always invertible for convenience.

The GREG estimator in equation (3) can be rewritten in calibration form as $t_{y_GREG} = \sum_S w_k y_k$, where

$$w_k = a_k + \left(T_{\mathbf{x}} - \sum_{j \in S} a_j \mathbf{x}_j \right) \left(\sum_{j \in S} c_j a_j \mathbf{x}'_j \mathbf{x}_j \right)^{-1} c_k a_k \mathbf{x}'_k.$$

Strictly speaking, the w_k are functions of the realized sample, S , and the $c_k a_k$, but we suppress that in the notation for convenience. Observe that the calibration weights can be expressed as

$$w_k = a_k (1 + c_k \mathbf{x}_k \mathbf{q}), \quad (4)$$

where $\mathbf{q} = [(\sum_S a_j c_j \mathbf{x}'_j \mathbf{x}_j)^{-1}]' (T_{\mathbf{x}} - \sum_S a_j \mathbf{x}_j)'$ is a column vector, since $\mathbf{x}_k \mathbf{q} = \mathbf{q}' \mathbf{x}'_k$.

Let us assume that reasonable regularity conditions hold (see, for example, Kott 2004a for a more thorough treatment) and the sample plan is such that $t_{y_E} - T_y = O_p(N/\sqrt{n})$, where n is the expected size of S (the actual size can be random), $\sum_S a_k \mathbf{x}_k - T_{\mathbf{x}} = \mathbf{O}_p(N/\sqrt{n})$, and

$\sum_S a_k c_k \mathbf{x}'_k \mathbf{f}_k - \sum_U c_k \mathbf{x}'_k \mathbf{f}_k = \mathbf{O}_p(N/\sqrt{n})$, where \mathbf{f}_k can be \mathbf{x}_k or y_k . Let $e_k = y_k - \mathbf{x}_k (\sum_U c_j \mathbf{x}'_j \mathbf{x}_j)^{-1} \sum_U c_j \mathbf{x}'_j y_j$, so that $\sum_U c_j \mathbf{x}'_j e_j = 0$, and $\sum_S a_k c_k \mathbf{x}'_k e_k = \mathbf{O}_p(N/\sqrt{n})$. We can express the error of t_{y_GREG} as

$$\begin{aligned} & t_{y_GREG} - T_y \\ &= \sum_{k \in S} w_k y_k - \sum_{k \in U} y_k \\ &= \sum_{k \in S} w_k e_k - \sum_{k \in U} e_k \left(\text{since } \sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k \right) \\ &= \sum_{k \in S} a_k e_k + \left(T_{\mathbf{x}} - \sum_{k \in S} a_k \mathbf{x}_k \right) \left(\sum_{k \in S} a_k c_k \mathbf{x}'_k \mathbf{x}_k \right)^{-1} \sum_{k \in S} a_k c_k \mathbf{x}'_k e_k \\ &\quad - \sum_{k \in U} e_k \\ &= \sum_{k \in S} a_k e_k - \sum_{k \in U} e_k + O_p(N/n). \end{aligned} \quad (5)$$

It is now not hard to see that the GREG estimator is randomization consistent; that is, $\text{plim}_{n \rightarrow \infty} [(t_{y_GREG} - T_y)/N] = 0$. Moreover, both the relative randomization bias and relative randomization mean squared error of the GREG estimator are order $1/n$. Since mean squared error = bias² + variance, we can conclude that the randomization bias of the GREG estimator is usually an asymptotically insignificant contributor to its mean squared error.

3. Redefining Calibration Weights

In their original definition of calibration weights, Deville and Särndal (1992) required that the set of calibration weights, $\{w_k | k \in S\}$ minimize some distance function between the members of the set and the original sampling weights, the a_k , subject to satisfying the calibration equation. As a result, the calibration estimator, $t_{y_CAL} = \sum_S w_k y_k$, was both unbiased under the model in equation (2) and usually randomization consistent.

Estevao and Särndal (2002) suggested removing the requirement that the calibration weights minimize a distance function. Instead, they essentially proposed that the w_k need only satisfy the calibration equation and be of the "functional form:"

$$w_k = a_k (1 + \mathbf{h}_k \mathbf{q}), \quad (6)$$

where \mathbf{h}_k is a row vector with the same dimension as \mathbf{x}_k such that $\sum_S a_k \mathbf{h}'_k \mathbf{x}_k$ is invertible, and \mathbf{q} is a column vector of that same dimension. Equation (6) is a mild generalization of (4) where \mathbf{h}_k effectively replaces $c_k \mathbf{x}_k$.

It is not hard to see that $\mathbf{q} = [(\sum_S a_j \mathbf{h}'_j \mathbf{x}_j)^{-1}]' (T_{\mathbf{x}} - \sum_S a_j \mathbf{x}_j)'$. Moreover, under mild conditions we assume to hold, $t_{y_CAL} = \sum_S w_k y_k = \sum_S a_k y_k + (T_{\mathbf{x}} - \sum_S a_j \mathbf{x}_j) (\sum_S a_j \mathbf{h}'_j \mathbf{x}_j)^{-1} \sum_S a_k \mathbf{h}'_k y_k$ is randomization consistent

whenever t_{y_E} is. It is unbiased under the linear prediction model in equation (2) when $E(\varepsilon_k | \{\mathbf{x}_k, \mathbf{h}_k | g \in S\}, \{I_g | g \in U\}) = 0$ for all $k \in U$.

This suggests another alternative definition of calibration weights: a set of weights, $\{w_k | k \in S\}$, such that,

- i. the w_k satisfy the calibration equation for $\{\mathbf{x}_k | k \in U\}$ and,
- ii. $t_{y_CAL} = \sum_S w_k y_k$ is randomization consistent whenever t_{y_E} is under mild conditions.

That is the definition we will use. This broadened definition of calibration weighting will prove very helpful when using calibration to adjust for nonresponse or coverage errors.

It follows from our new definition that Estevao and Särndal's functional-form calibration is indeed a form a calibration weighting. Borrowing from econometric theory, the components of \mathbf{h}_k that are not linear combinations of components of \mathbf{x}_k are called "instrumental variables."

4. Possibly Nonlinear Calibration

Building on ideas in Deville and Särndal (1992), we can generalize the linear form for the calibration weights in equation (6) to

$$w_{k_GEN} = a_k f(\mathbf{h}_k, \mathbf{q}^*), \quad (7)$$

where f is a monotonic, twice-differentiable function with $f(0)=1, f'(0)=1$ ($f'(0)$ is the first derivative of f evaluated at 0), and \mathbf{q}^* is chosen so that the calibration equation holds. Unlike the calibration-weight equation above, the calibration equation itself, $\sum_S w_k \mathbf{x}_k = T_{\mathbf{x}}$, remains linear. Note that since $f(0)=1, f'(0)=1, f(\mathbf{h}_k, \mathbf{q}^*) \approx 1 + \mathbf{h}_k \mathbf{q}^*$.

Strictly speaking, there should be an additional symbol on w_{k_GEN} (and later on w_{k_LIN}) to denote the particular choice of \mathbf{h}_k . It has been dropped for convenience.

A solution, \mathbf{q}^* , to equation (7) can often be reached iteratively. One can start with $\mathbf{q}^{(0)} = \mathbf{0}$; that is, $\sum_S w_k^{(0)} y_k$, where $w_k^{(0)} = a_k f(0)$. For $r=1, 2, \dots$, one then sets $\mathbf{q}^{(r)} = \mathbf{q}^{(r-1)} + \{[\sum_S f'(\mathbf{h}_k, \mathbf{q}^{(r-1)}) a_k \mathbf{x}_k' \mathbf{h}_k]^{-1} (T_{\mathbf{x}} - \sum_S w_k^{(r-1)} \mathbf{x}_k)'\}$, and $w_k^{(r)} = a_k f(\mathbf{h}_k, \mathbf{q}^{(r)})$. Iteration stops at r^* when $T_{\mathbf{x}} = \sum_S w_k^{(r^*)} \mathbf{x}_k$ for all practical purposes. One should be aware, however, that there may not be a set of weights that can be expressed in the form of equation (7) while satisfying the calibration equation.

Note that $\mathbf{q}^{(1)}$ above equals the \mathbf{q} in $w_{k_LIN} = a_k (1 + \mathbf{h}_k \mathbf{q})$. A Taylor expansion around zero reveals $f(\mathbf{h}_k, \mathbf{q}^{(1)}) = 1 + \mathbf{h}_k \mathbf{q}^{(1)} + O_p(1/n)$ under mild conditions, so $\sum_S w_k^{(1)} y_k = \sum_S w_{k_LIN} y_k + O_p(N/n) = T_y [1 + O_p(1/n)]$.

Furthermore, it is not difficult to see that $w_{k_GEN} = w_{k_LIN} [1 + O_p(1/n)]$, an equality that proves helpful in variance estimation.

The most common example in practice of a nonlinear f is $f(\mathbf{h}_k, \mathbf{q}) = \exp(\mathbf{x}_k \mathbf{q})$, where the values of each of the components of \mathbf{x}_k , denoted x_{1k}, \dots, x_{pk} , are either 0 or 1. That is effectively the form of Deming and Stephan's (1940) raking weights computed via iterative proportional fitting. Many have observed that the iterative routine described above can be used even when the components of \mathbf{x}_k are not binary as they are in Deming and Stephan. Note that the generalized raking calibration weights that result are always nonnegative.

5. Variance Estimation

Särndal, Swensson, and Wretman (1989) proposed this *plug-in* model variance/randomization mean-squared-error estimator for t_{y_GREG} under an arbitrary sampling plan:

$$v_{SSW} = \sum_{k \in S} \sum_{j \in S} [(\pi_{kj} - \pi_k \pi_j) / \pi_{kj}] (w_k r_k) (w_j r_j). \quad (8)$$

The term derives from r_k being "plugged into" v_{SSW} in place of the unknown $e_k = y_k - \mathbf{x}_k (\sum_U \mathbf{h}_i' \mathbf{x}_i)^{-1} \sum_U \mathbf{h}_i' y_i$ for randomization-mean-squared-error estimation.

Paralleling arguments in Deville and Särndal (1992), v_{SSW} also applies more generally to t_{y_CAL} with calibration weights defined by equation (7) with

$$r_k = y_k - \mathbf{x}_k \left(\sum_{j \in S} a_j \mathbf{h}_j' \mathbf{x}_j \right)^{-1} \sum_{j \in S} a_j \mathbf{h}_j' y_j. \quad (9)$$

This is because $w_{k_GEN} = w_{k_LIN} [1 + O_p(1/n)]$, so $\sum_S w_{k_GEN} e_k = \sum_S w_{k_LIN} e_k + O_p(N/n) = \sum_S a_k e_k + O_p(N/n)$. The last step uses reasoning exhibited in equation (5) with \mathbf{h}_j serving in place of the $c_j \mathbf{x}_j$.

In their article, Deville and Särndal effectively replace the a_j in equation (9) with $w_j = a_j f(\mathbf{h}_j, \mathbf{q}^*)$. A different version is given in Demanti and Rao (2004), where the a_j in the equation are replaced by $a_j f'(\mathbf{h}_j, \mathbf{q}^*)$. This author noted in a comment accompanying the latter that all three versions of the r_k are asymptotically identical since $f(0) = f'(0) = 1$ and \mathbf{q}^* is asymptotically $\mathbf{0}$. These asymptotic identities may no longer hold when calibration weighting is used to adjust for nonresponse as we shall see in the following section.

Developing asymptotic properties for v_{SSW} under stratified simple random sampling is a simple matter. In this context, v_{SSW} collapses to

$$v_{ST1} = \sum_{\alpha=1}^A (n_{\alpha} / [n_{\alpha} - 1]) \sum_{k \in S_{\alpha}} (1 - n_{\alpha} / N_{\alpha}) \times \left(w_k r_k - \sum_{j \in S_{\alpha}} w_j r_j / n_{\alpha} \right)^2,$$

where S_α denotes the sample of n_α units in stratum α ($\alpha = 1, \dots, A$), and U_α the stratum population containing N_α elements.

For a multi-stage sample it makes sense to allow the possibility that ε_k and ε_j in the prediction model are correlated when k and j are elements in the same PSU, but not otherwise. When finite-population correction can be ignored, the model variance of a calibration estimator is approximately $V_m = \sum_{i \in S'} E_e [(\sum_{k \in S(i)} w_k \varepsilon_k)^2]$ under mild conditions, where $S(i)$ is the set of sampled elements in PSU i , and S' is the set of PSUs selected in the first stage of sampling.

The following variance estimator, not strictly equal to v_{SSW} , often has good randomization and model-based properties (when the first-stage selection probabilities are all small):

$$v_{ST2} = \sum_{\alpha=1}^A (n_\alpha / [n_\alpha - 1]) \times \left\{ \sum_{j \in S_\alpha} - \left(\sum_{k \in S_{\alpha j}} w_k r_k \right)^2 \frac{\left(\sum_{j \in S_\alpha} \sum_{k \in S_{\alpha j}} w_k r_k \right)^2}{n_\alpha} \right\}, \quad (10)$$

where α denotes a first-stage stratum of PSU's, n_α the number of sampled PSU's in stratum α , S_α the set of sampled PSU's in α , and $S_{\alpha j}$ the set of subsampled elements from PSU j of stratum α . There can be many stages of sampling involved.

It is not hard to show that v_{ST2} is asymptotically indistinguishable from the jackknife variance estimator:

$$v_J = \sum_{\alpha=1}^A ([n_\alpha - 1] / n_\alpha) \left\{ \sum_{j \in S_\alpha} (t_{y_CAL(\alpha j)} - t_{y_CAL})^2 \right\}, \quad (11)$$

where $t_{y_CAL(\alpha j)} = \sum_{k \in S} w_{k(\alpha j)} y_k$, and the *jackknife replicate calibration weights* are

$$w_{k(\alpha j)} = w_k a_{k(\alpha j)} / a_k + \left(\sum_{m \in U} \mathbf{x}_m - \sum_{m \in S} w_m [a_{m(\alpha j)} / a_m] \mathbf{x}_m \right) \times \left(\sum_{m \in S} a_{m(\alpha j)} \mathbf{h}'_m \mathbf{x}_m \right)^{-1} a_{k(\alpha j)} \mathbf{h}'_k, \quad (12)$$

where $a_{k(\alpha j)} = 0$ when k is in PSU j of stratum α , $a_{k(\alpha j)} = a_k$ when k is not in stratum α at all, and $a_{k(\alpha j)} = (n_\alpha / [n_\alpha - 1]) a_k$ otherwise. The $w_{k(\alpha j)}$ are constrained so that $\sum_{k \in S} w_{k(\alpha j)} \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$ for all αj .

Let $S(\alpha+)$ be the set of *elements* in stratum α (not PSU's like S_α), and $S(\alpha j)$ the set of elements in PSU j of stratum α . Under mild conditions we assume to hold,

$$\begin{aligned} \sum_U \mathbf{x}_m - \sum_S w_m [a_{m(\alpha j)} / a_m] \mathbf{x}_m \\ = (n_\alpha / [n_\alpha - 1]) \left(\sum_{S(\alpha j)} w_k \mathbf{x}_k - \sum_{S(\alpha+)} w_k \mathbf{x}_k / n_\alpha \right) = \mathbf{O}_P(N/n), \\ \sum_S a_{m(\alpha j)} \mathbf{h}'_m \mathbf{x}_m = \mathbf{O}_P(N), \\ \text{and } \sum_S a_{m(\alpha j)} \mathbf{h}'_m e_m = \mathbf{O}_P(N/\sqrt{n}). \end{aligned}$$

As a result,

$$\begin{aligned} t_{y_CAL} - t_{y_CAL} = \sum_S w_{k(\alpha j)} e_k - \sum_S w_k e_k \\ = (n_\alpha / [n_\alpha - 1]) \left(\sum_{S(\alpha+)} w_k e_k / n_\alpha - \sum_{S(\alpha j)} w_k e_k \right) \\ + O_P(N/n^{3/2}), \end{aligned}$$

and $v_J = v_{ST2} [1 + O_P(1/\sqrt{n})]$ when $p \lim_{n \rightarrow \infty} (n v_{ST2} / N^2) > 0$.

The replicate weights defined in equation (12) do not require iteration even when the calibration weights are themselves produced that way. This is a great computation convenience. It not only saves computer time, it avoids the possibility that an iterative solution for the w_k may exist while one for the replicate weights does not.

6. Unit Nonresponse

6.1 Quasi-randomization and Prediction Modeling

In this section we explore handling unit (whole-element) nonresponse as an additional phase of Poisson sampling. That is the essence of a *quasi-randomization* model. Each element k in the original sample, now denoted F , is assumed to have a probability of response, p_k . The probability of elements k and j jointly responding is $p_k p_j$, and whether element k would respond (given a vector of covariates) is independent of whether it is chosen for the original sample.

It is often possible to construct a set of weights so that the calibration estimator is randomization consistent under the quasi-randomization model. We are interested here in a particular way of constructing those weights. To this end, we assume that the quasi-randomization model is correct. Each element has attached to it a row vector of auxiliary variables, \mathbf{x}_k , for which $T_x = \sum_U \mathbf{x}_j$ is known. Finally, each p_k is assumed to have the form:

$$p_k = 1 / f(\mathbf{h}_k \boldsymbol{\phi}), \quad (13)$$

where $\boldsymbol{\phi}$ is an unknown column vector, \mathbf{h}_k is a row vector with the same dimension as \mathbf{x}_k , and $\sum_S a_k \mathbf{h}'_k \mathbf{x}_k / N$, where S now denotes the "subsample" of respondents, is invertible both for the realized population size, N , and in the probability limit.

The function $f(\cdot)$ in equation (13) is assumed to be monotonic and twice differentiable. Its functional form is known, but the value of the governing parameter, $\boldsymbol{\phi}$, is not. When plugged into the calibration-weight equation,

$w_k = a_k f(\mathbf{h}_k \mathbf{q})$, so that the calibration equation itself, $\sum_s w_k \mathbf{x}_k = T_{\mathbf{x}}$ holds, $f(\mathbf{h}_k \mathbf{q})$ implicitly estimates the inverse of the element response probabilities. Unlike when calibration is used to correct for $\sum_s a_k \mathbf{x}$ differing from $T_{\mathbf{x}}$ due purely to sampling error, $f(0)$ and $f'(0)$ do not need to be 1 nor does $\mathbf{h}_k \phi$ need to be zero.

The most obvious choice for \mathbf{h}_k when postulating the response model in equation (13) is \mathbf{x}_k itself. In a common example of calibration weighting for nonresponse, the components of \mathbf{x}_k are indicator variables: $x_{gk} = 1$ when k is in group g and zero otherwise. When the groups are mutually exclusive, calibration weighting is the same thing as reweighting within post-stratification classes. See, for example, Särndal, Swensson and Wretman (1992, page 585). The prediction model usually underpinning calibration (the prefix "prediction" is needed to distinguish this model from the quasi-randomization one) assumes that every element k in group g , whether or not it would respond, has a common mean: $E(y_k) = \beta_g$. The quasi-random response model is analogous: $p_k = 1/\phi_g$. The two models are conceptually very different, however.

When the groups are not mutually exclusive, raking is one method of determining calibration weights. There are others depending on the exact form of the assumed response function $f(\cdot)$. The prediction model remains linear, $E(y_k) = \mathbf{x}_k \beta$, while the response model that leads to raking, $p_k = \exp\{-\mathbf{x}_k \phi\}$, does not. Berry, Flatt, and Pierce (1996) provides an example of using raking to adjust for nonresponse.

In many applications of calibration weighting the components of \mathbf{x}_k are continuous or semi-continuous rather than dichotomous. In an annual crop survey, for example, let x_{1k} be the quantity of corn harvested in the previous census of agriculture by farm k , x_{2k} be the farm's harvested wheat, x_{3k} its harvested potatoes, and so forth. The annual crop survey has an assumed prediction model for farm k 's planted corn acres, y_{1k} , of the form: $y_{1k} = \mathbf{x}_k \beta_{1k} + \varepsilon_{1k}$. The subscript, 1, is corn-specific. There are other survey values of interest, like planted wheat acres, and potentially assumed prediction models for each.

The quasi-random response model for the crop survey depends on assumptions about $f(\cdot)$ and \mathbf{h}_k in equation (13) with \mathbf{h}_k possibly equal to \mathbf{x}_k . Unlike the prediction model, the same assumed quasi-randomization model applies for all survey variables.

Promising choices for $f(\cdot)$ are $\exp(\cdot)$ and $1 + \exp(\cdot)$, the latter corresponding to a response probabilities being fit by a logistic function of $\mathbf{h}_k \phi$. It may also be reasonable to assume $h_{gk} = x_{gk}^\lambda$ for $\lambda < 1$. In particular, setting $\lambda = 0$ means that the probability of farm k responding to the annual crop survey depends only on whether the farm had

corn, wheat, or potatoes on the previous census of agriculture rather than on how much of those crops it had.

In the crop-survey example, the components of \mathbf{x}_k from the previous census were the best predictors available for the corresponding annual survey values *before* sampling. Whether farm k responds to the survey, however, is more likely a function of the farm's current planted corn acres, if any, than on a predetermined proxy for that value. As a result, placing survey values in \mathbf{h}_k rather than corresponding census values is tempting. There is a theoretical problem with this procedure as we shall see.

Given an $f(\cdot)$, the iterative method described in Section 4 will often be able to uncover a row vector \mathbf{q} such that $T_{\mathbf{x}} = \sum_s a_k f(\mathbf{h}_k \mathbf{q}) \mathbf{x}_k$. When that happens, estimating T_y with $t_{y_CAL} = \sum_s w_k y_k$, where $w_k = a_k f(\mathbf{h}_k \mathbf{q})$, will have good properties under the linear prediction model: $y_k = \mathbf{x}_k \beta + \varepsilon_k$, where $E(\varepsilon_k | \{\mathbf{x}_k, \mathbf{h}_k, I_g | g \in U\}) = 0$ for all $k \in U$, $I_k = 1$ if element k is both in the original sample and responds, 0 otherwise.

Prediction-model unbiasedness is simply a result of the weights satisfying the calibration equation. Note, however, that if components of \mathbf{h}_k come from the survey rather than \mathbf{x}_k , the prediction-model assumption that $E(\varepsilon_k | \mathbf{h}_k) = 0$ can be problematic. At the extreme, consider the case where one such component is y_k itself. Usually, $E(\varepsilon_k | y_k)$ is not 0. In the crop-survey example described earlier, y_k can be the annual corn acres planted on farm k . Putting this value in \mathbf{h}_k makes the associated calibration estimator for corn prediction-model biased.

When the prediction model is correct (treating $E(\varepsilon_k | \{\mathbf{x}_k, \mathbf{h}_k, I_g | g \in U\}) = 0$ as an integral part of the model), however, calibration weighting based on any choice of $f(\cdot)$ will produce estimators with good prediction-model-based properties. These estimators will also have good quasi-randomization properties when the response model in equation (13) is correct for that choice of $f(\cdot)$. In some sense, one model provides protection against the failure of the other. See Kott (1994).

As noted, the prediction model is more likely to hold when $\mathbf{h}_g = \mathbf{x}_g$. Even then, sometimes the ε_k in the model in equation (2) satisfy $E(\varepsilon_k | \{\mathbf{x}_g | g \in U\}) = 0$, but not $E(\varepsilon_k | \{\mathbf{x}_g, I_g | g \in U\}) = 0$; that is to say, the sampling mechanism – including response – is not ignorable with respect to the prediction model.

We can factor I_k into $I_{k1} I_{k2}$, where $I_{k1} = 1$ if and only if k is in the original sample, and $I_{k2} = 1$ if and only if k would respond if sampled. The interested reader can confirm that calibration weighting provides some protection against bias if the prediction model in equation (2) holds when $E(\varepsilon_k | \{\mathbf{x}_g, \mathbf{h}_g, I_{g2} | g \in U\}) = 0$; that is when the response mechanism is ignorable with respect to the

prediction model but not necessarily the original sampling mechanism.

6.2 Quasi-randomization Mean Squared Error Estimation

Whether or not t_{y_CAL} can reasonably be called prediction-model unbiased has no effect on its quasi-randomization-based properties. Note that $\mathbf{h}_k\phi$ are $\mathbf{h}_k\mathbf{q}$ are scalar values not vectors. Since $T_{\mathbf{x}} = \sum_S a_k f(\mathbf{h}_k\mathbf{q})\mathbf{x}_k$, our assumptions and the mean value theorem ($f(\mathbf{h}_k\phi) = f(\mathbf{h}_k\mathbf{q}) + f'(\theta_k)(\mathbf{h}_k\phi - \mathbf{h}_k\mathbf{q})$) reveal

$$\begin{aligned} T_{\mathbf{x}} - \sum_{k \in S} a_k f(\mathbf{h}_k\phi)\mathbf{x}_k &= \sum_{k \in S} a_k [f'(\theta_k)\mathbf{h}_k(\mathbf{q} - \phi)]\mathbf{x}_k \\ &= \mathbf{O}_p(N/\sqrt{n}) \end{aligned}$$

for some scalar θ_k between each $\mathbf{h}_k\mathbf{q}$ and $\mathbf{h}_k\phi$. From this we see that if $\sum_S a_j f'(\mathbf{h}_j\phi)\mathbf{h}'_j\mathbf{x}_j/N$ is invertible both for the realized N and at the probability limit (recall that f is monotonic so f' is never zero), then

$$\begin{aligned} \mathbf{q} - \phi &= \left\{ \left[\sum_{j \in S} a_j f'(\mathbf{h}_j\mathbf{q})\mathbf{h}'_j\mathbf{x}_j \right]^{-1} \right\}' \left[T_{\mathbf{x}} - \sum_{i \in S} a_i f(\mathbf{h}_i\phi)\mathbf{x}_i \right] \\ &= \mathbf{O}_p(1/\sqrt{n}) \\ &= \left\{ \left[\sum_{j \in S} a_j f'(\mathbf{h}_j\phi)\mathbf{h}'_j\mathbf{x}_j \right]^{-1} \right\}' \left[T_{\mathbf{x}} - \sum_{i \in S} a_i f(\mathbf{h}_i\phi)\mathbf{x}_i \right] \\ &\quad + \mathbf{O}_p(1/n). \end{aligned}$$

The estimator t_{y_CAL} has an error of

$$\begin{aligned} t_{y_CAL} - T_y &= \sum_{k \in S} a_k f(\mathbf{h}_k\mathbf{q})y_k - \sum_{k \in U} y_k \\ &= \sum_{k \in S} a_k f(\mathbf{h}_k\mathbf{q})e_k - \sum_{k \in U} e_k, \end{aligned}$$

where

$$e_k = y_k - \mathbf{x}_k \left(\sum_{U'} f'(\mathbf{h}_j\phi)p_j\mathbf{h}'_j\mathbf{x}_j \right)^{-1} \sum_{U'} f'(\mathbf{h}_j\phi)p_j\mathbf{h}'_j\mathbf{y}_j,$$

and $p_j = 1/f(\mathbf{h}_j\phi)$. The e_k are again unknown. They have been design so that $\sum_S a_k f'(\mathbf{h}_k\phi)\mathbf{h}'_k e_k = \mathbf{O}_p(N/\sqrt{n})$. Continuing:

$$\begin{aligned} t_{y_CAL} - T_y &= \sum_{k \in S} a_k f(\mathbf{h}_k\phi)e_k - \sum_{k \in U} e_k + \sum_{k \in S} a_k \{f(\mathbf{h}_k\mathbf{q}) - f(\mathbf{h}_k\phi)\}e_k \\ &= \sum_{k \in S} a_k f(\mathbf{h}_k\phi)e_k - \sum_{k \in U} e_k + \sum_{k \in S} a_k f'(\mathbf{h}_k\phi)\mathbf{h}_k(\mathbf{q} - \phi)e_k \\ &\quad + O_p(N/n) \\ &= \sum_{k \in S} a_k f(\mathbf{h}_k\phi)e_k - \sum_{k \in U} e_k + (\mathbf{q} - \phi)' \sum_{k \in S} a_k f'(\mathbf{h}_k\phi)\mathbf{h}'_k e_k \\ &\quad + O_p(N/n) \\ &= \sum_{k \in S} a_k f(\mathbf{h}_k\phi)e_k - \sum_{k \in U} e_k + O_p(N/n). \end{aligned} \quad (14)$$

Thus, t_{y_CAL} is quasi-randomization consistent under mild conditions whenever $t = \sum_S a_k f(\mathbf{h}_k\phi)y_k$ is.

To estimate the quasi-randomization mean squared error of t_{y_CAL} (i.e., the estimator's randomization mean squared error under the response model), we first note that the

probability that elements k and j , $k \neq j$, are both in the respondent subsample is $\pi_{kj}^* = \pi_{kj}p_kp_j$. Let $\pi_k^* = \pi_kp_k$, and recall that $a_k = 1/\pi_k^*$ and $1/p_k = f(\mathbf{h}_k\phi)$. From equation (14), we see that the quasi-randomization mean squared error of t_{y_CAL} is approximately

$$\begin{aligned} E_1[(t_{y_CAL} - T_y)^2] &\approx \sum_{k \in U} \sum_{j \in U} (\pi_{kj}^* - \pi_k^*\pi_j^*)(e_k/\pi_k^*)(e_j/\pi_j^*) \\ &= \sum_{k \in U} (1 - \pi_k^*)e_k^2/\pi_k^* \\ &\quad + \sum_{k \in U} \sum_{j \in U, k \neq j} (\pi_{kj} - \pi_k\pi_j)(e_k/\pi_k)(e_j/\pi_j). \end{aligned} \quad (15)$$

If the original sample is Poisson, then $v_m = \sum_S (w_k^2 - w_k)r_k^2$ with

$$r_k = y_k - \mathbf{x}_k \left[\sum_{j \in S} a_j f'(\mathbf{h}_j\mathbf{q})\mathbf{h}'_j\mathbf{x}_j \right]^{-1} \sum_{j \in S} a_j f'(\mathbf{h}_j\mathbf{q})\mathbf{h}'_j\mathbf{y}_j, \quad (16)$$

serves as both a reasonable estimator for prediction-model variance and quasi-randomization mean squared error under mild conditions, since $w_k \approx 1/\pi_k^*$ and $r_k \approx e_k$. A close relative of the non-intuitive sample residual in equation (16) can be found in Folsom and Singh (2000). See Kott (2004a) for a further discussion of v_m in a purely sampling context.

For a general design, we can get close to a good variance/mean-squared-error estimator with

$$\begin{aligned} v_{\text{com}} &= \sum_{k \in S} (w_k^2 - w_k)r_k^2 \\ &\quad + \sum_{k \in S} \sum_{j \in S, k \neq j} [(\pi_{kj} - \pi_k\pi_j)/\pi_{kj}](w_kr_k)(w_jr_j). \end{aligned} \quad (17)$$

The right hand side of equation (17) estimates the right hand side of equation (15) with r_k replacing e_k . Note that $\sum_U (1 - \pi_k^*)e_k^2/\pi_k^*$ in equation (15) is estimated by $\sum_S (w_k^2 - w_k)r_k^2$ rather than $\sum_S w_k^2(1 - \pi_k^*)r_k^2$, which would make v_{com} more consistent with v_{SSW} in equation (8). This substitution results in a variance estimator with good prediction-model-based properties when the ε_k are uncorrelated, and $\sigma_k^2 = \mathbf{x}_k\zeta$, for some ζ . It can be made even in the absence of nonresponse.

When the actual sample is multistage, and the first stage selection probabilities are ignorably small, v_{ST2} in equation (10) can be used as the variance/mean-squared-error estimator with r_k defined once more by equation (16).

When f is linear, $f'(\theta) = 1$, and the r_k in equation (16) are computed as if there were no nonresponse. The same holds true for the variance/mean-squared-error estimator v_{ST2} . Unfortunately, this f corresponds to an awkward-looking response-probability function: $p_k = 1/\mathbf{h}_k\phi$. Fuller, Loughin and Baker (1994) made these observations for the case where $\mathbf{h}_k = c_k\mathbf{x}_k$.

The jackknife, v_J , in equation (11) can be computed with these jackknife replicate weights:

$$w_{k(\alpha_j)} = w_k a_{k(\alpha_j)} / a_k + \left(\sum_{m \in U} \mathbf{x}_m - \sum_{m \in S} w_m [a_{m(\alpha_j)} / a_m] \mathbf{x}_m \right) \times \left(\sum_{m \in S} a_{m(\alpha_j)} f'(\mathbf{h}_m \mathbf{q}) \mathbf{h}_m' \mathbf{x}_m \right)^{-1} a_{k(\alpha_j)} f'(\mathbf{h}_k \mathbf{q}) \mathbf{h}_k', \quad (18)$$

an obvious generalization of the jackknife replicate weights in equation (12). Again when $f'(\theta) = 1$, v_J can be computed as if there were no nonresponse.

7. Coverage Modeling

Folsom and Singh (2000) pointed out that the treatment of nonresponse through calibration weighting can also be used to adjust for undercoverage. In the context, the quasi-random phase as sampling occurs conceptually before the actual sample is drawn. The population associated with the sampling frame is assumed to be a Poisson sample from a hypothetical complete population for which the vector T_k must be known. The frame population becomes F , while the hypothetical complete population is U . The probability that element $k \in U$ is in F is assumed to be modeled correctly by equation (13). If the first (from U to F) and second (from F to S) phases of sampling are independent, then all the theory developed for using calibration weighting to handle nonresponse carries over to handling undercoverage.

It should be noted that coverage adjustment through calibration is an extension of the well-known practice of coverage adjustment through post-stratification often used with telephone surveys. As with the post-stratification special case, one needs quantities for the calibration targets for U that can be assumed to be free of error or to have very little mean squared error compared to the calibration estimators themselves.

Folsom and Singh noted that overcoverage (duplication) or a combination of under and overcoverage can be handled with their methodology. The definition of p_k in equation (13) becomes the expected number of times k is in the frame, which can now exceed 1 due to potential duplication.

Folsom and Singh further suggested that $f(\cdot)$ have the flexible form:

$$f(\mathbf{x}_k \phi) = \frac{U(C - L) \exp(\mathbf{x}_k \phi) + L(U - C)}{(U - C) + (C - L) \exp(\mathbf{x}_k \phi)}, \quad (19)$$

where $L \geq 0$, $1 < U \leq \infty$, and $L < C \leq U$ are predetermined constants. They call this the "General Exponential Model" or "GEM." Observe that when $C = 1$, $U = \infty$, and $L = 0$, $p_k = 1/f(\mathbf{x}_k \phi) = \exp(-\mathbf{x}_k \phi)$. Similarly, when $C = 2$, $U = \infty$, and $L = 1$, $p_k = [1 + \exp(\mathbf{x}_k \phi)]^{-1}$; that is to say, the probability of coverage (or response) is logistic. The values

L and U serve as bounds on the calibration adjustment, $f(\cdot)$, while $C = f(0)$ is effectively its center.

The authors made the calibration adjustment in GEM even more flexible by postulating three classes of sampling units, each with its own set of U , C , and L values. They proposed its use both for coverage-error and unit-nonresponse adjustment

8. A Small Empirical Example

Since the jackknife replicate weights expressed in equation (18) are new, it is prudent to investigate whether they actually work with real data. To this end, the author took the MU281 data from Särndal, Swensson and Wretman (1992) and replicated it 20 times (so $N = 5,620$). Using stratified simple random sampling, 16 units were selected from each of the eight unequally-sized strata. The variable RMT85 served as y_k and P75 as x_k in $\mathbf{x}_k = (1, x_k)$. Each of the 128 sampled units was given a probability of being in the respondent subsample, S , which decreased with the size of x_k ; in particular, $p_k = \exp(-0.35 x_k / M_x)$, where M_x was the population mean of the x_k . In 1,600 simulations, the size of the S ranged from 78 to 110, with an average of approximately 93.8.

The total T_y was estimated two ways, with $t_{y_LIN} = \sum_S a_k (1 + \mathbf{x}_k \mathbf{q}) y_k$ and with $t_{y_EXP} = \sum_S a_k \exp(\mathbf{x}_k \mathbf{q}^{(EXP)}) y_k$, where \mathbf{q} and $\mathbf{q}^{(EXP)}$ were respectively selected so that the calibration equation held. The former was a GREG estimator, while the latter was a generalized raking estimator. Both estimators were unbiased under the implied prediction model ($y_k = \mathbf{x}_k \beta + \varepsilon_k$), but only t_{y_EXP} was randomization consistent under the correct response model. The GREG implicitly assumed $p_k = 1/(\phi_0^{(LIN)} + \phi_1^{(LIN)} \mathbf{x}_k)$ for unknown $\phi_0^{(LIN)}$ and $\phi_1^{(LIN)}$.

The small size of the sample relative to the population in each stratum allowed the ignoring of finite population correction in variance/mean-squared-error estimation (called "variance estimation" from now on). Variances were estimated using i , the linearization estimator, v_{ST2} , in equation (10) with r_k defined by equation (16), and ii , the proposed jackknife, v_J , in equation (11) with replicate weights defined by equation (18). To make the jackknife computations easier, the 16 samples in each stratum were randomly assigned to one of four clusters, so that only 32 jackknife replicates had to be computed.

For comparison purposes, a better version of the linearization variance estimator, labeled $v_{ST2(e)}$, was also computed with r_k replaced by $e_k = y_k - \mathbf{x}_k (\sum_U f'(\mathbf{x}_j \phi) p_j \mathbf{x}_j')^{-1} \sum_U f'(\mathbf{x}_j \phi) p_j \mathbf{x}_j' y_j$, where ϕ and p_j were known. In practice, e_k is rarely known, but computing $v_{ST2(e)}$ is useful here for comparison.

One should note that computations of r_k and e_k were slightly different depending on whether the variance estimator for t_{y_LIN} or for t_{y_EXP} was of interest. For t_{y_LIN} , $f'(\mathbf{x}_j \boldsymbol{\phi}) = f'(\mathbf{x}_j \mathbf{q}) = 1$; for t_{y_EXP} , $f'(\mathbf{x}_j \mathbf{q}^{(exp)}) = \exp(\mathbf{x}_j \mathbf{q}^{(exp)})$, and $f'(\mathbf{x}_j \boldsymbol{\phi}) = 1/p_j$.

Table 1 displays the empirical means (the mean over the 1,600 simulations) of the two estimators for T_y normalized so that $T_y = 100$. Although both are close to unbiased, t_{y_LIN} is significantly different from 100 at the 0.05 level; t_{y_EXP} is not. This is not surprising, since only the latter is based on the correct response model.

The variance estimators and empirical mean squared errors of each estimator were normalized so that the empirical means of the respective $v_{ST2(e)}$'s were 100. Neither $v_{ST2(e)}$ had an empirical mean significantly different from the empirical mean squared error (EMSE) of the associated estimator. This was a bit disappointing. It seems that although t_{y_LIN} had a significant empirical bias, this bias was such a small component of the estimator's mean squared error, that the difference between its EMSE and the empirical mean of $v_{ST2(e)}$ was not significant.

The $v_{ST2(e)}$ were chosen as benchmarks for the table rather than the empirical mean squared errors because each $v_{ST2(e)}$ had roughly half the empirical standard error of the corresponding EMSE (which itself was the average of 1,600 squared differences) and correlated more strongly with the variance estimators. The t -values for this part of the table were also computed with respect to the $v_{ST2(e)}$.

The two linearization variance estimators had surprisingly large downward biases. Apparently, there was a tendency for unusually large w_{k_LIN} and w_{k_EXP} to cause associated r_k to be appreciably smaller than e_k in absolute terms. The problems associated with unusually large w_{k_LIN} and w_{k_EXP} seem to be more muted with the jackknives.

To speed up the asymptotics of the linearization variance estimators (*i.e.*, reduce the difference between r_k and e_k), an *ad-hoc* adjustment of v_{ST2} was computed by replacing each r_k with $r_k^{(adjusted)} = r_k / \omega_k$, where $\omega_k^2 = 1 - \mathbf{x}_k (\sum_j a_j f'(\mathbf{x}_j \mathbf{q}) \mathbf{x}_j' \mathbf{x}_j)^{-1} a_k f'(\mathbf{x}_k \mathbf{q}) \mathbf{x}_k' = 1 + O_p(1/n)$. Observe that under the prediction model with the ε_k uncorrelated and $E(\varepsilon_k^2) = \sigma_k^2$, $E(r_k^{(adjusted)}) \approx \sigma_k^2$. The near equality is exact when all the $a_j f'(\mathbf{x}_j \mathbf{q})$ and σ_j , respectively, are equal.

The adjusted v_{ST2} for both t_{y_LIN} and t_{y_EXP} remained biased downward, while the v_j were biased upward by a slightly smaller amount. Although these biases were significant, they were reasonably small (from 4.5 to 11.2%) and suggest that the variance estimators may have indeed been asymptotically unbiased as theoretically demonstrated in previous sections.

Using $v_{ST2(e)}$ as an efficient proxy for EMSE, the empirical mean squared error of t_{y_EXP} , which incorporated the correct response model, was more than 13% larger than that of the t_{y_LIN} , which did not. One should not generalize broadly based on one data set involving only two calibration variables, however. See Crouse and Kott (2004) for a different set of results.

Table 1
Empirical Means of Estimators Based on 1,600 Simulations*

Empirical mean (standard error)		t – value (two-sided significance)	
The Estimators for $T_y(T_y = 100)$			
t_{y_LIN}	99.84 (0.06)	–2.79 (0.02)	difference from T_y
t_{y_EXP}	100.04 (0.06)	0.58 (0.56)	
Variance Estimators for $t_{y_LIN}(E_{EMP}(v_{ST2(e)}) = 100)$			
v_{ST2}	83.59 (1.53)	–19.96 (< 0.0001)	difference from $v_{ST2(e)}$
$v_{ST2(adjusted)}$	95.53 (1.80)	–6.09 (< 0.0001)	
v_J	104.69 (2.28)	3.60 (0.0003)	
EMSE	99.35 –	–0.18 (0.85)	
Variance Estimators for $t_{y_EXP}(E_{EMP}(v_{ST2(e)}) = 100)$			
v_{ST2}	73.12 (1.54)	–18.22 (< 0.0001)	difference from $v_{ST2(e)}$
$v_{ST2(adjusted)}$	88.79 (1.98)	–8.57 (< 0.0001)	
v_J	107.00 (2.73)	4.09 (< 0.0001)	
EMSE	101.21 –	0.33 (0.74)	
Other Statistics			
relvar ($v_{ST2(e)[LIN]}$)	0.051 –	–	
relvar ($v_{ST2(e)[EXP]}$)	0.059 –	–	
$\frac{(v_{ST2(e)[LIN]} - v_{ST2(e)[EXP]})}{(E_{EMP}(v_{ST2(e)[EXP]}))}$	–0.1340 (0.010)	–13.87 (< 0.0001)	

* In four additional simulations, convergence was not reached in 10 iterations for t_{y_EXP} . They were excluded from the analysis.

Whether or not one is better off incorporating the correct response model in the calibration estimator, if one does so, then the variance estimators discussed in the previous section, perhaps with the linearization estimator adjusted as suggested in this section, appear to be serviceable.

A second set of 1,600 simulations (not displayed) were done using the same population and stratified sampling design but with each sampled element given a 70% chance of being in the respondent sample (the average respondent sample size was roughly 89.8). In this set of simulations, both estimators for T_y are randomization consistent under the response model. Consequently, it is not surprising, that the empirical means of t_{y_LIN} and t_{y_EXP} were virtually identical (within 0.01% of each other) as were their empirical mean squared errors (within 1% of each other). The empirical means of each pair of variance estimators (e.g., var_{ST2} for t_{y_LIN} and t_{y_EXP}) were likewise very close (within 1% of each other). The relative bias of the adjusted v_{ST2} (compared to $\text{var}_{ST2(e)}$) was -1.3% when estimating the variance of t_{y_LIN} and -2.2% when estimating the variance of t_{y_EXP} . The relative biases of the unadjusted linearization variances were -9.0% and -10.3%, respectively. The relative bias of both jackknives was 3.6%.

9. Discussion

9.1 Estimating a Response Model Explicitly

When faced with unit nonresponse, many have attempted to estimate the element probabilities of response, $p_k = 1/f(\mathbf{h}_k, \phi)$, directly. This method requires one to have information on \mathbf{h}_k for every element in the sample whether it responded to the survey or not, but \mathbf{h}_k need not have the same dimension as \mathbf{x}_k . The direct-adjustment method is generally not available for handling coverage errors.

Fuller (2002) noted that there can be an extra term in the quasi-randomization mean squared error of $t_{y_GREG} = \sum_S a_k^* y_k + (T_x - \sum_S a_j^* x_j) (\sum_S c_j a_j^* x_j^* x_j)^{-1} \sum_S c_k a_k^* x_k^* x_k$, where S is the respondent subsample, $a_k^* = a_k [1 + f(\mathbf{h}_k, \mathbf{q})]$, and \mathbf{q} is a consistent direct estimator for the quasi-randomization model parameter, ϕ . This does not imply that direct estimation of the response model based on a given $f(\cdot)$ and \mathbf{h}_k is less efficient than analogous calibration when \mathbf{h}_k has the same dimension of \mathbf{x}_k . See Kim (2004) for a suggestion otherwise. Nevertheless, the convenience of incorporating nonresponse adjustment into calibration is appealing when variance estimates need to be produced.

A reasonable compromise is to choose the form of $f(\cdot)$ and \mathbf{h}_k by modeling the response behavior of the entire sample and then estimating the parameter of $f(\cdot)$ implicitly through calibration. This compromise also overcomes a striking weakness of using calibration weighting to adjust

for nonresponse (as well as for coverage errors). The choices for $f(\cdot)$ and \mathbf{h}_k are motivated primarily by plausibility and convenience and not by a statistical analysis of the data.

9.2 Response Homogeneity Groups

To control the magnitude of the weight adjustment due to nonresponse, Little (1986) recommended that one estimate \mathbf{q} explicitly and then divide the sample into C mutually exclusive groups based on the sizes of the fitted $f(\mathbf{h}_k, \mathbf{q})$ values. One then computes the adjusted weight for each element k in group c as with post-stratification: $w_{k_ADJ} = (\sum_{F(c)} w_g / \sum_{S(c)} w_g) w_k$, where $F(c)$ is that part of the original sample in group c , $S(c)$ is the subsample of $F(c)$ that respond, and w_k is the sampling weight assigned to element k after sampling but before quasi-random subsampling. This approach assumes that each element in a group has (roughly) the same probability of response, hence the term "response homogeneity group."

An alternative way of incorporating fitted $f(\mathbf{h}_k, \mathbf{q})$ values into the estimation based on methodology developed in the text follows. Divide the fitted values into P groups based in their sizes, where P is again the dimension of \mathbf{x}_k , and let \mathbf{d}_k be a row vector of indicator variables for the P cells. By setting each $w_k = a_k [1 + (T_x - \sum_S a_j x_j) \times (\sum_S a_j \mathbf{d}_j' x_j)^{-1} \mathbf{d}_k']$, one computes a set of weights for the respondent subsample that, unlike $\{w_{k_ADJ}\}$ above, satisfies the calibration equation for the respondent sample. Because of the nature of \mathbf{d}_k , this linear method returns the same set of calibration weights as fitting $w_k = a_k \exp(\mathbf{d}_k \mathbf{f})$ would – if both produce a set of weights. Note that since calibration weights can be negative with the linear method, it may be able to find a set that the generalized raking method cannot. The linear method effectively scales the a_k -value for every element in the same group by a fixed amount. Thus, it may not produce surprisingly small or surprisingly large weights when the dimension of \mathbf{x}_k is small compared to the sample size.

9.3 Breaking Up Sample and Nonresponse Calibration

In the previous section we noted that it is possible for components of \mathbf{h}_k in equation (13), the quasi-random response model, to be unknown before enumeration. When such an \mathbf{h}_k is used in calibration, it might no longer to reasonable to assert that the resulting t_{y_CAL} is prediction-model unbiased. This is particularly troublesome when nonresponse is modest compared to the sample size. An intriguing idea is to calibrate in two phases. The first phase, sample calibration, adjusts for the difference between T_x and $\sum_F a_k x_k$, and would not include any components in \mathbf{h}_k unavailable at the time of sampling. The second phase,

nonresponse calibration, adjusts for the difference between $\sum_F a_k \mathbf{x}_k$ and $\sum_S a_k \mathbf{x}_k$ and would include component variables only available after the respondent subsample is enumerated.

A more thorough analysis of this idea must wait for another time.

9.4 Work at NASS

The National Agricultural Statistics Service (NASS) used variants of the Fuller *et al.* (1994) approach for handling undercoverage in the 2002 Census of Agriculture (see Fetter and Kott 2003) and for adjusting an agricultural economics survey with large nonresponse to match totals from more reliable surveys (see Crouse and Kott 2004). In this approach, $f(\cdot)$ has the form:

$$f(\mathbf{x}_k \phi) = \begin{cases} L & \text{when } \mathbf{x}_k \phi < L \\ \mathbf{x}_k \phi & \text{when } L \leq \mathbf{x}_k \phi \leq U \\ U & \text{when } \mathbf{x}_k \phi > U, \end{cases} \quad (20)$$

which truncates linear calibration at pre-specified values, L and U , to control the size of the weight adjustment. Note that when $f(\cdot) = U$ or L , $f'(\cdot) = 0$. Unlike the calibration adjustment in equation (19), $f(\cdot)$ in equation (20) is not twice differentiable at L or U . This does not cause a problem in practice.

The agency's original justification for calibration in these contexts was based on prediction-modeling. Equation (20) is simple to implement and appears to produce weights within an acceptable range more often than readily available alternatives.

NASS is investigating the following questions: How sensitive is t_{v_CAL} to the choice of $f(\cdot)$ in practice? Would a different choice for $f(\cdot)$ result in less bias, and if so, would the reduction in absolute bias translate into a lower mean squared error? What would be the effect of replacing some component of the vector of calibration variables with a better predictor of nonresponse/undercoverage?

References

- Berry, C.C., Flatt, S.W. and Pierce, J.P. (1996). Correcting unit nonresponse via nonresponse modeling and raking in the California Tobacco Survey. *Journal of Official Statistics*, 12, 349-363.
- Crouse, C., and Kott, P.S. (2004). Evaluation alternative calibration schemes for an economic survey with large nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- Deming, W.E., and Stephan, F.F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal total are known. *Annals of Mathematical Statistics*, 11, 427-444.
- Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimation for survey data. *Survey Methodology*, 30, 17-26.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Estevao, V.M., and Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16, 379-399.
- Fetter, M.J., and Kott, P.S. (2003). Developing a coverage adjustment strategy for the 2002 Census of Agriculture. Presented at 2003 Federal Committee on Statistical Methodology Research Conference, http://www.fcsm.gov/03papers/fetter_kott.pdf.
- Folsom, R.E., and Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 598-603.
- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.
- Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting for the 1987-88 National Food Consumption Survey. *Survey Methodology*, 20, 75-85.
- Kim, J.K. (2004). Efficient nonresponse weighting adjustment using estimated response probability. *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- Kott, P.S. (1990). The design consistent regression estimator and its conditional variance. *Journal of Statistical Planning and Inference*, 24, 287-296.
- Kott, P.S. (1994). A note on handling nonresponse in surveys. *Journal of the American Statistical Association*, 89, 693-696.
- Kott, P.S. (2004a). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 48, 263-277.
- Kott, P.S. (2004b). Comment. *Survey Methodology*, 30, 27-28.
- Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- Lundström, S., and Särndal, C.-E. (1999). Calibration as a standard method for the treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of a finite population total. *Biometrika*, 76, 527-537.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data

Jerome P. Reiter, Trivellore E. Raghunathan and Satkartar K. Kinney¹

Abstract

The theory of multiple imputation for missing data requires that imputations be made conditional on the sampling design. However, most standard software packages for performing model-based multiple imputation assume simple random samples, leading many practitioners not to account for complex sample design features, such as stratification and clustering, in their imputations. Theory predicts that analyses of such multiply-imputed data sets can yield biased estimates from the design-based perspective. In this article, we illustrate through simulation that (i) the bias can be severe when the design features are related to the survey variables of interest, and (ii) the bias can be reduced by controlling for the design features in the imputation models. The simulations also illustrate that conditioning on irrelevant design features in the imputation models can yield conservative inferences, provided that the models include other relevant predictors. These results suggest a prescription for imputers: the safest course of action is to include design variables in the specification of imputation models. Using real data, we demonstrate a simple approach for incorporating complex design features that can be used with some of the standard software packages for creating multiple imputations.

Key Words: Complex sampling design; Multiple imputation; Nonresponse; Surveys.

1. Introduction

Typically in large surveys, less than 100% of the sampled units respond fully to the survey. Some units do not respond at all, and others respond only to certain items. One approach to handle such nonresponse is multiple imputation of missing data (Rubin 1987). It has been used in, for example, the Fatality Analysis Reporting System (Heitjan and Little 1991), the Consumer Expenditures Survey (Raghunathan and Paulin 1998), the National Health and Nutrition Examination Survey (Schafer, Ezzati-Rice, Johnson, Khare, Little and Rubin 1998), the Survey of Consumer Finances (Kennickell 1998) and the National Health Interview Survey (Schenker, Raghunathan, Chiu, Makuc, Zhang and Cohen 2005). Multiple imputation also has been suggested to protect confidentiality of public-release data (Rubin 1993; Little 1993; Raghunathan, Reiter and Rubin 2003; Reiter 2003, 2004, 2005). See Rubin (1996) and Barnard and Meng (1999) for a review of other applications.

Multiple imputation, in theory, conditions on the sampling design when deriving methods for obtaining inferences from multiply-imputed datasets (Rubin 1987). However, imputers seldom account for complex sampling design features, such as stratification and clustering, when using available software packages to construct imputation models. They instead use multivariate normal or general location models (*e.g.*, the software NORM written by Joe Schafer), or use sequential regression models (Raghunathan,

Lepkowschi, van Hoewyk and Solenberger 2001). These methods can be modified to incorporate design features, but this is infrequently done.

This paper has two objectives. First, we illustrate the bias that can arise when imputers fail to account for complex design features in imputation models. To do so, we simulate multiple imputation in two-stage, stratified-cluster samples. The simulations indicate these biases can be severe, even when using design-based estimators in multiply-imputed data sets with moderate amounts of missing data. Second, we suggest two simple approaches to account for design features in imputation models. The first approach, which is relatively easy to implement, includes dummy variables for stratum or cluster effects in the imputation models. The second approach, which is computationally more complex than the first, uses hierarchical models where (i) the effects of clustering are incorporated using random effects, and (ii) the effects of stratification are incorporated using fixed effects. The simulations show that accounting for the design in these ways can reduce the bias. They also illustrate that controlling for design features that are unrelated to the survey variables can result in inefficient, but conservative, inferences relative to those from models that do not condition on such features, provided that the models include the predictors required to make the missing at random assumption (Rubin 1976) plausible. We demonstrate the first approach to incorporating the design features by imputing missing data from the National Health and Nutrition Examination Survey using a sequential regression approach.

1. Jerome P. Reiter and Satkartar K. Kinney, Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708, U.S.A.; Trivellore E. Raghunathan, Department of Biostatistics and Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, U.S.A.

2. Inferences from Multiply-Imputed Data Sets

To describe construction of and inferences from multiply-imputed data sets, we use the notation of Rubin (1987). For a finite population of size N , let $I_j = 1$ if unit j is selected in the original survey, and $I_j = 0$ otherwise, where $j = 1, 2, \dots, N$. Let $I = (I_1, \dots, I_N)$. Let n be the size of the sample obtained using a complex design. To simplify notation, assume only one variable in the survey is subject to nonresponse. Let $R_j = 1$ if unit j responds to the original survey, and $R_j = 0$ otherwise. Let $R = (R_1, \dots, R_N)$. The notation can be extended to handle multivariate item nonresponse, but such complication is not necessary for our purposes.

Let Y be the $N \times p$ matrix of survey data for all units in the population. Let $Y_{\text{inc}} = (Y_{\text{obs}}, Y_{\text{mis}})$ be the $n \times p$ matrix of survey data for units with $I_j = 1$; Y_{obs} is the portion of Y_{inc} that is observed, and Y_{mis} is the portion of Y_{inc} that is missing due to nonresponse. Let Z be the $N \times d$ matrix of design variables for all N units in the population, e.g., stratum or cluster indicators or size measures. We assume that such design information is known at least approximately, for example from census records or the sampling frames.

Values for Y_{mis} are usually constructed from draws from some approximation to the Bayesian posterior predictive distribution of $(Y_{\text{mis}} | Z, Y_{\text{obs}}, I, R)$. These draws are repeated independently $l = 1, \dots, M$ times to obtain M completed data sets, $D^{(l)} = (Z, Y_{\text{obs}}, Y_{\text{mis}}^{(l)}, I, R)$.

From these multiply-imputed data sets, some user of the data seeks inferences about some estimand $Q = Q(Z, Y)$. For example, Q could be a population mean or a population regression coefficient. In each imputed data set $D^{(l)}$, the analyst estimates Q with some estimator q and the variance of q with some estimator u . We assume that the analyst specifies q and u by acting as if each $D^{(l)}$ was in fact collected data from a random sample of (Z, Y) based on the original sampling design I , i.e., q and u are complete-data estimators.

For $l = 1, \dots, M$, let $q^{(l)}$ and $u^{(l)}$ be respectively the values of q and u in data set $D^{(l)}$. Under assumptions described in (Rubin 1987), the analyst can obtain valid inferences for scalar Q by combining the $q^{(l)}$ and $u^{(l)}$. Specifically, the following quantities are needed for inferences:

$$\bar{q}_M = \sum_{l=1}^M q^{(l)} / M \quad (1)$$

$$b_M = \sum_{l=1}^M (q^{(l)} - \bar{q}_M)^2 / (M - 1) \quad (2)$$

$$\bar{u}_M = \sum_{l=1}^M u^{(l)} / M. \quad (3)$$

The analyst then can use \bar{q}_M to estimate Q and $T_M = (1 + \frac{1}{M})b_M + \bar{u}_M$ to estimate the variance of \bar{q}_M . When n and M are large, inferences for scalar Q can be based on normal distributions, so that a $(1 - \alpha)\%$ confidence interval for Q is $\bar{q}_M \pm z(\alpha/2)\sqrt{T_M}$. For moderate M , inferences can be based on t -distributions with degrees of freedom $\nu_M = (M - 1)(1 + r_M^{-1})^2$, where $r_M = (1 + M^{-1})b_M / \bar{u}_M$, so that a $(1 - \alpha)\%$ confidence interval for Q is $\bar{q}_M \pm t_{\nu_M}(\alpha/2)\sqrt{T_M}$. Refinements of these basic combining rules have been proposed by several authors, including Li, Raghunathan and Rubin (1991a), Li, Meng and Rubin (1991b), Raghunathan and Siscovick (1996), and Barnard and Rubin (1999).

3. Illustrative Simulations

In this section, we use simulations to illustrate the biases/inefficiencies associated with incorporating design features in imputation models. We simulate three target populations of $N = 100,000$ units that are stratified and clustered within strata. In the first population, Y depends on both stratum and cluster effects. In the second population, Y depends on strata but not on cluster effects. In the third population, Y is unrelated to the stratum and cluster indicators. The first population is used to demonstrate the importance of including all relevant design variables, and the second and third populations are used to examine the effect of including irrelevant design variables. The simulated populations are stylized to illustrate the importance of modeling the survey design; hence, the magnitudes of biases/inefficiencies may not be generalizable to other settings.

Each population is divided into five equally-sized strata comprised of $N_h = 200$ clusters, for $h = 1, \dots, 5$. Each cluster c in stratum h is comprised of N_{hc} units. In each stratum, ten clusters have $N_{hc} = 300$, twenty clusters have $N_{hc} = 200$, sixty clusters have $N_{hc} = 100$, sixty clusters have $N_{hc} = 75$, and fifty clusters have $N_{hc} = 50$. Cluster sizes are varied to magnify design effects when taking multi-stage cluster samples. For each target population, there are two survey variables, X and Y . In all three populations, for simplicity we generate each X_{hcj} , where j indexes a unit within stratum and cluster hc , from $X_{hcj} \sim N(0, 10^2)$. To generate Y , we use different methods for each population, as shall be described in subsequent sections.

We randomly sample units from each population using multi-stage cluster sampling. First, we take a simple random sample of $n_1 = 40$ clusters from stratum 1, $n_2 = 20$ clusters from stratum 2, $n_3 = 30$ clusters from stratum 3, $n_4 = 10$ clusters from stratum 4, and $n_5 = 15$ clusters from stratum 5. The cluster sample sizes differ across strata to magnify

design effects relative to equal sampling. We then take a simple random sample of twenty units from each sampled cluster. Hence, there are 2,300 units with $I_{hcl} = 1$.

The estimands of interest in each population are $Q = \bar{Y}$, the population mean of Y , and the coefficients for the population regression of Y on X . The complete-data estimator of \bar{Y} is the usual, unbiased design-based estimator,

$$q = \frac{1}{100,000} \left(\sum_{h=1}^5 \frac{200}{n_h} \sum_{c \in h} \hat{y}_{hc} \right),$$

where $\hat{y}_{hc} = N_{hc} \bar{y}_{hc}$ is the estimated total in cluster hc . The complete-data estimator of the variance of q is,

$$u = \frac{1}{100,000^2} \left(\sum_{h=1}^5 200^2 \left(1 - \frac{n_h}{200} \right) s_h^2 / n_h + \sum_{h=1}^5 \frac{200}{n_h} \sum_{c \in h} N_{hc}^2 \left(1 - \frac{20}{N_{hc}} \right) s_{hc}^2 / 20 \right),$$

where s_h^2 is the sample variance of the \hat{y}_{hc} and s_{hc}^2 is the sample variance of Y within cluster hc . The estimators of the coefficients in the regression of Y on X are the usual approximately unbiased, design-based estimators, which are computed using the “survey” routines (Lumley 2004) in the software package R. These routines estimate variances using Taylor series linearizations. These estimators are used for all multiply-imputed data sets in all simulations.

For each sample, we let X be fully observed, and let Y be missing for about 30% of the sampled units.

Each unit’s binary response variable, R_{hcj} , is drawn from a Bernoulli distribution:

$$\Pr(R_{hcj} = 1) = \frac{\exp(-0.847 - 0.1 X_{hcj})}{1 + \exp(-0.847 - 0.1 X_{hcj})} \tag{4}$$

Here, $R_{hcj} = 1$ means that the unit’s value of Y is missing. Equation 4 implies that Y_{mis} is missing at random (Rubin 1976). We can ignore the missing data mechanism provided that imputations for missing data are conditional on X . We purposefully do not allow missingness to depend on stratum or cluster membership to illustrate that bias can arise from failing to account for the survey design even when the ignorable missing data mechanism does not depend on the sampling design. Of course, if the sampling design is related to missingness, as it is in many real datasets, one must condition on the sampling design to make the missing data mechanism ignorable.

We examine three strategies to impute Y_{mis} that make different use of the design information. These strategies are summarized in Table 1. The first strategy, labeled SRS, completely disregards the sampling design. The second strategy, FX, incorporates the stratification and the

clustering by using fixed effects for each cluster within stratum. The third strategy, HM, uses normal random effects models that incorporate the stratification and clustering. For SRS, one model is fit to the entire data set. For FX and HM, models are fit separately in each stratum. All three strategies regress on X because it is part of the missing data mechanism; not conditioning on X would violate ignorability and cause bias.

Table 1
Imputation Strategies

Label	Imputation model for missing Y_{hcj}
SRS	$N(\beta_0 + \beta_1 X_{hcj}, \sigma^2)$
FX	$N(\beta_{0h} + \beta_{1h} X_{hcj} + \omega_{hc}, \sigma_h^2)$
HM	$N(\beta_{0h} + \beta_{1h} X_{hcl} + \omega_{hc}, \sigma_h^2), \omega_{hc} \sim N(0, \tau^2)$

All imputations are draws from the appropriate Bayesian posterior predictive distributions. First, we draw parameters of the imputation models from their posterior distributions given the components of the observed data, (Z, X, Y_{obs}, I, R) , that are included in the models. Second, we draw values of the missing data from the distributions given in Table 1. Diffuse priors are used for all parameters. For strategy HM, we draw values of the parameters using a Gibbs sampler (Gelfand and Smith 1990). We run the sampler for a burn-in period to get approximate convergence, then we use every tenth draw for imputations. Finally, we use $M = 5$ independently drawn imputations in each data set for each strategy.

3.1 Simulation A: Illustration of Disregarding Relevant Design Features

In this simulation, we generate a population in which the distributions of Y differ across strata and clusters. We call this “Population 1”. Specifically, for unit j in stratum h and cluster c , we construct the population value of Y_{hcj} from

$$Y_{hcj} = 10 X_{hcj} + \beta_{0h} + \omega_{hc} + \epsilon_{hcj} \tag{5}$$

where β_{0h} is a scalar constant for stratum h , the ω_{hc} is a scalar constant for cluster hc , and ϵ_{hcj} is a random error term drawn from $N(0, 200^2)$. The values of the stratum effects are $\beta_{01} = 500, \beta_{02} = -250, \beta_{03} = 0, \beta_{04} = 250,$ and $\beta_{05} = -500$. The values of the ω_{hc} are obtained by drawing five sets of $N_h = 200$ values from independent $N(0, 70^2)$. The stratum and cluster effects are widely dispersed to magnify design effects relative to simple random sampling, which in turn magnifies the effects of disregarding the design in imputations. We then sample from this population using the stratified cluster sampling scheme outlined previously. We create the missing data indicator R using equation 4.

Table 2 shows the results of 1,000 replications of the three imputation strategies outlined in Table 1. The additional row labeled “Complete data” shows the results using the data for all sampled units, *i.e.*, assuming no units with $I_{h_{cj}} = 1$ have $R_{h_{cj}} = 0$. The column labeled “95% CI cov.” contains the percentage of the 1,000 simulated confidence intervals that contain the population parameter. The column labeled “Pt. Est.” contains the averages of the 1,000 point estimates of Q . The column labeled “Var” contains the variances of the 1,000 point estimates of Q . The column labeled “Est. Var” contains the averages across the 1,000 replications of the estimated variances of the point estimates. The columns labeled “Var(Est.Var)” and “MSE(Est.Var)” give the variance and mean squared error of the 1,000 estimated variances.

Imputations based on method SRS lead to severely biased estimates and very poor confidence interval coverage in this population. These problems exist even though there is not much missing information and despite the fact that we use design-unbiased estimators for inferences. Both FX and HM have point estimates that approximately match the complete-data point estimates, and both have coverage rates that approximately match the rates for the complete data inferences. FX and HM have similar profiles because the fixed effect models and the hierarchical models produce similar estimates of the parameters in equation 5.

When estimating the population mean, the variance associated with FX or HM is only slightly larger than the variance associated with the complete-data estimator. This is because of the large cluster effects, which makes the within-imputation variance a dominant factor relative to the between-imputation variance. That is, the fraction of missing information due to missing data is relatively small when compared to the effect of clustering.

3.2 Simulation B: Illustration of including irrelevant predictors

Modeling the design features is essential when the features are related to the survey variables of interest. How does modeling irrelevant design features affect inferences? In this section, we present the results of two simulation studies that explore this question.

First, we generate “Population 2” in which the distribution of Y differs across strata but does not depend on the clusters. To do so, we use the same generation method as in Equation 5, setting the ω_{hc} equal to zero. The $\epsilon_{h_{cj}}$ are drawn from $N(0, 100^2)$. We sample from Population 2 and generate missing data using the schemes outlined previously. The results for 1,000 replications are displayed in Table 3.

SRS continues to have severe bias and poor confidence interval coverage because it ignores the stratification. For FX and HM, the averages of their point estimates are within simulation error of the average of the point estimates for the complete data. Additionally, their confidence interval coverage rates approximately match the coverage rate for the complete-data intervals. This indicates that FX and HM are reasonable for these populations, even though the irrelevant cluster features are included in their imputation models.

We next generate “Population” 3 in which the distribution of Y is independent of the strata and cluster membership indicators. Specifically, to generate Y , we subtract the β_{0h} from the values of Y generated in Population 2. We then sample from Population 3 using the stratified cluster sampling scheme and create missing data using the methods outlined previously. The results for 1,000 replications are displayed in Table 4.

Table 2
Performance of Imputation Procedures when the Design Features are Related to the Survey Variable of Interest.
The Population Mean Equals 3.2 and the Population Regression Coefficients Equal 3.0 and 10.1

	Method	95% CI cov.	Pt. Est.	Var	Est. Var	Var(Est. Var)	MSE (Est. Var)
Mean Y	Complete data	94.2	2.0	544.91	527.31	31,626.19	31,936.07
	SRS	38.0	45.8	327.79	360.74	11,927.97	13,013.35
	FX	94.8	2.4	554.09	579.92	37,474.82	38,141.70
	HM	94.5	2.3	551.02	553.16	34,056.39	34,060.99
Intercept	Complete data	93.0	2.4	529.51	499.73	18,543.13	19,430.21
	SRS	39.5	46.8	340.09	365.50	9,351.15	9,996.99
	FX	94.5	2.8	539.19	551.68	21,529.16	21,685.33
	HM	93.9	2.7	536.82	524.82	19,256.24	19,400.11
Slope	Complete data	93.3	10.1	1.24	1.15	0.14	0.15
	SRS	64.8	7.6	2.10	2.20	0.55	0.56
	FX	94.5	10.1	1.45	1.44	0.18	0.18
	HM	95.7	10.1	1.53	1.65	0.29	0.30

Table 3

Performance of Imputation Procedures when the Population has Stratum Effects but no Cluster Effects.
The Population Mean Equals 0.34 and the Population Regression Coefficients Equal 0.14 and 10.13

	Method	95% CI cov.	Pt. Est.	Var	Est. Var	Var(Est. Var)	MSE (Est. Var)
Mean Y	Complete data	93.6	0.37	468.97	461.88	29,301.77	29,352.04
	SRS	31.1	42.90	259.46	303.46	10,228.40	12,164.74
	FX	93.7	0.32	473.86	474.21	30,408.95	30,409.07
	HM	93.4	0.34	476.03	465.53	29,406.61	29,516.85
Intercept	Complete data	93.0	0.72	451.46	432.74	14,955.20	15,305.73
	SRS	31.5	43.10	275.22	311.36	8,134.04	9,440.57
	FX	93.2	0.66	456.08	444.88	15,539.21	15,664.64
	HM	92.3	0.68	457.48	436.25	14,941.00	15,391.75
Slope	Complete data	93.1	10.09	0.99	0.91	0.09	0.10
	SRS	59.0	7.72	1.67	1.77	0.35	0.36
	FX	93.4	10.10	1.03	0.98	0.10	0.10
	HM	93.3	10.10	1.03	0.96	0.10	0.10

Table 4

Performance of Imputation Procedures when the Design Variables are Completely Unrelated to the Survey Variable of Interest.
The Population Mean Equals 0.34 and the Population Regression Coefficients Equal 0.14 and 10.04

	Method	95% CI cov.	Pt. Est.	Var	Est. Var	Var(Est. Var)	MSE (Est. Var)
Mean Y	Complete data	94.7	0.35	14.61	14.73	32.65	32.66
	SRS	95.7	0.12	16.45	19.22	40.65	48.31
	FX	97.8	0.40	19.64	28.29	97.66	172.38
	HM	95.1	0.26	18.77	19.16	47.29	47.44
Intercept	Complete data	93.7	0.12	7.13	7.20	5.31	5.32
	SRS	96.8	-0.10	8.97	11.72	13.59	21.10
	FX	98.6	0.17	12.23	20.62	39.84	110.24
	HM	96.2	0.03	10.45	11.61	15.09	16.45
Slope	Complete data	94.5	10.04	0.07	0.07	0.001	0.001
	SRS	96.4	10.07	0.10	0.13	0.002	0.003
	FX	96.4	10.04	0.12	0.15	0.003	0.004
	HM	95.2	10.05	0.11	0.12	0.002	0.002

SRS finally produces point estimates whose averages are within simulation error of the complete data average point estimate. This is because the imputations in SRS reflect the population structure reasonably well. This suggests that disregarding the design in imputation models may provide acceptable inferences when the design variables are only weakly correlated with the survey outcomes. As in the previous simulations, FX and HM continue to have average point estimates within simulation error of the complete-data average point estimate. When comparing the three imputation strategies, we see that FX and HM are inefficient relative to SRS. This is because the imputation models for FX and HM estimate parameters that equal approximately zero in the population, whereas SRS sets them equal to zero. HM has smaller variance than FX does, because the hierarchical imputation model smoothes the estimated cluster effects towards zero.

For FX, the percentage of confidence intervals that cover Q is larger than the percentages for the complete-data intervals and HM intervals. This is because the estimated variance for FX tends to be larger than its actual variance.

This apparent upward bias in T_M also exists for SRS, resulting in a larger coverage percentage than those for the complete-data and HM.

4. Real Data Example

We next examine the effect of accounting for stratification and clustering when imputing missing data in a genuine dataset. The data are taken from the public use file for the 1999–2002 National Health and Nutrition Examination Surveys. Individuals are grouped in 56 clusters divided among 28 strata. Many variables have 5% to 10% missing data.

We imputed missing data using two strategies: one ignoring design variables (like SRS) and one incorporating the design variables using fixed effects for cluster indicators (like FX). In the imputation model, we included 27 dummy variables to represent 28 strata and one dummy variable within-each stratum to represent the two clusters nested within each stratum. That is, a total of 55 dummy variables

were included as predictors. We used a stepwise variable selection procedure to identify statistically significant interactions between these dummy variables and survey variables, and we included these interactions as predictors in the imputation model as well. The values were imputed using the sequential regression approach implemented in the software package IVEWARE (www.isr.umich.edu/src/smp/ive). We generate $M=10$ data sets for each strategy.

We consider three estimands. The first is the population percentage of people who have ever had their blood cholesterol level checked (BPQ060). This variable has about 15% missing values. The second and third are the population regression coefficients in a logistic regression of BPQ060 on family poverty income ratio (INDFMPIR), a continuous variable that has about 12% missing values. These estimands are estimated using design-based methods computed with the “survey” routines in the software package R.

Table 5 displays the results for both imputation strategies. The two sets of estimates for all analyses are very similar. In this case, incorporating the design variables into the imputation model hardly impacts the results. This is due in part to the small fractions of missing information and the relative unimportance of stratum and cluster effects. However, there is minimal penalty for including the design features in the imputation model. In light of the results of the simulations in section 3, we would incorporate the design features in this imputation model.

Table 5
Comparison of Real Data Results when Design Features
are Included in Imputation Model and when
Design Features are Ignored

	Pt. Est.	S.E.	95% CI
Mean BPQ060			
design	0.319	0.010	(0.299, 0.339)
no design	0.319	0.011	(0.296, 0.341)
Intercept: Logistic Regression			
design	0.362	0.054	(0.256, 0.467)
no design	0.352	0.052	(0.251, 0.454)
Slope: Logistic Regression			
design	-0.409	0.020	(-0.449, -0.369)
no design	-0.407	0.019	(-0.444, -0.371)

5. Concluding Remarks

The simulation studies, though limited, suggest disregarding the sampling design in multiple imputation can be a risky practice. When the design variables are related to the survey variables, as in our Simulation A, failing to include the design variables can lead to severe bias. On the other

hand, including irrelevant design variables, as in our Simulation B and the NHANES example, leads at worst to inefficient and conservative inferences when the imputation models are otherwise properly specified.

Including dummy variables for cluster effects greatly reduced the bias relative to disregarding the design completely. However, blindly including dummy variables is not an automatic solution. When the regression slopes or variances differ across clusters, using FX or HM may result in biased estimates, since important design features are disregarded. Imputers suspecting such relationships should include appropriate interactions with the dummy variables for the design features, as we did in the NHANES example. In some surveys the design may be so complicated that it is impractical to include dummy variables for every cluster. In these cases, imputers can simplify the model for the design variables, for example collapsing cluster categories or including proxy variables (e.g., cluster size) that are related to the outcome of interest.

The simulations suggest that there can be payoffs to using hierarchical models for imputation of missing data relative to using fixed effects models, particularly when cluster effects are similar. However, hierarchical models are more difficult to fit than fixed effect models. For example, it is daunting to fit hierarchical models in complex designs when data are missing for several continuous and categorical variables. It may be possible to fit sequential hierarchical models in a spirit similar to the sequential regression imputations of Raghunathan *et al.* (2001). This is an area for future research. A further disadvantage of hierarchical models is that they are easier to mis-specify than fixed effects models. For example, if the cluster effects follow a non-normal distribution, the hierarchical normal model used in this paper could provide implausible imputations.

With multiple imputation, the key to success is specifying an imputation model that reasonably describes the conditional distribution of the missing values given the observed values. Design features frequently are related to survey variables, so that including them in the imputation models reduces the risks of model mis-specification. We believe that in many cases the potential biases resulting from excluding important design variables, or other variables related to the missing data mechanism, outweigh the potential inefficiencies from estimating small coefficients. This reinforces the general advice provided by many on multiple imputation: include all variables that are related to the missing data in imputation models to make the missing data mechanism ignorable (e.g., Meng 1994; Little and Raghunathan 1997; Schafer 1997, and Collins, Schafer and Kam 2001).

Acknowledgements

This research was funded by the National Science Foundation grant ITR-0427889. The authors thank the associate editor and reviewers for their comments and suggestions.

References

- Barnard, J., and Meng, X. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8, 17-36.
- Barnard, J., and Rubin, D.B. (1999). Small-sample degrees of freedom with multiple-imputation. *Biometrika*, 86, 948-955.
- Collins, L.M., Schafer, J.L. and Kam, C.K. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, 6, 330-351.
- Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- Heitjan, D.F., and Little, R.J.A. (1991). Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics*, 40, 13-29.
- Kennickell, A.B. (1998). Multiple imputation in survey of consumer finances. In *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 11-20.
- Li, K.H., Raghunathan, T.E. and Rubin, D.B. (1991a). Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86, 1065-1073.
- Li, K.H., R.T.E., Meng, X.L. and Rubin, D.B. (1991b). Significance levels from repeated p -values with multiply-imputed data. *Statistica Sinica*, 1, 65-92.
- Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.
- Little, R.J.A., and Raghunathan, T.E. (1997). Should imputation of missing data condition on all observed variables? In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 617-622.
- Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9, 8.
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science*, 9, 538-558.
- Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*, 27, 85-96.
- Raghunathan, T.E., and Paulin, G.S. (1998). Multiple imputation of income in the Consumer Expenditure Survey: Evaluation of statistical inference. In *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 1-10.
- Raghunathan, T.E., Reiter, J.P. and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.
- Raghunathan, T.E., and Siscovick, D.S. (1996). A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45, 335-352.
- Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181-189.
- Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30, 235-242.
- Reiter, J.P. (2005). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 168, 185-205.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462-468.
- Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.
- Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.
- Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A. and Rubin, D.B. (1998). The NHANES III multiple imputation project. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 28-37.
- Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G. and Cohen, A.J. Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association*, forthcoming.

Bernoulli Bootstrap for Stratified Multistage Sampling

Fumio Funaoka, Hiroshi Saigo, Randy R. Sitter and Tsutom Toida¹

Abstract

In this article, we propose a Bernoulli-type bootstrap method that can easily handle multi-stage stratified designs where sampling fractions are large, provided simple random sampling without replacement is used at each stage. The method provides a set of replicate weights which yield consistent variance estimates for both smooth and non-smooth estimators. The method's strength is in its simplicity. It can easily be extended to any number of stages without much complication. The main idea is to either keep or replace a sampling unit at each stage with preassigned probabilities, to construct the bootstrap sample. A limited simulation study is presented to evaluate performance and, as an illustration, we apply the method to the 1997 Japanese National Survey of Prices.

Key Words: Complex survey; Linearization; Quantiles; Resampling; Stratification.

1. Introduction

Many large scale surveys are conducted using a stratified multi-stage sampling design. Variance estimation in this type of design can be analytically involved or even impossible. In addition, for publicly released data sets the specific forms of estimators the end-user may wish to obtain variance estimates for are unknown. As a result, resampling methods are often carried out to obtain a set of replicate weights that can be supplied with the data set and used for the purpose of variance estimation for a broad class of possible estimators. The bootstrap is particularly useful since it can handle both smooth and nonsmooth sample statistics under multistage designs. A concise summary of several bootstrap methods for finite population sampling is found in Shao and Tu (1995, pages 232-282) (see also, Gross 1980; Bickel and Freedman 1984; McCarthy and Snowden 1985; Rao and Wu 1988; Kovar, Rao and Wu 1988; Sitter 1992a, b; Booth, Butler and Hall 1994; Shao and Sitter 1996).

If the first-stage sampling fraction is small, there are various bootstrap methods available that treat the first-stage sampling as having been with-replacement for the purposes of variance estimation. In the case where the first-stage sampling fraction is not negligible, there are fewer results available. For bootstrapping in two-stage sampling with simple random sampling (SRS) at each stage see Sitter (1992a, 1992b) and with unequal probabilities Rao and Wu (1988). However, if the first-stage sampling fractions are not negligible no simple bootstrap procedure is available for three or more stages of sampling. In this paper, we propose a new bootstrap method which easily accommodates such cases when the sampling is simple random sampling (SRS)

at each stage. We call it a Bernoulli bootstrap (BBE) because of its resemblance to Bernoulli sampling. The National Survey of Prices (NSP) in Japan is used for illustration.

The paper is organized as follows. Section 2 introduces notation for three-stage stratified sampling. Section 3 describes two types of BBE. Section 4 investigates properties of the methods via simulation. Section 5 describes the sampling design of the 1997 NSP and illustrates the use of BBE on the NSP data. Concluding remarks are made in section 6.

2. Stratified Three-Stage Sampling

In stratified random sampling, the finite population, consisting of N primary sampling units (PSU's), is partitioned into H nonoverlapping strata of N_1, N_2, \dots, N_H PSU's, respectively; thus, $\sum_{h=1}^H N_h = N$. A simple random sample without replacement (SRSWOR) of PSU's is taken independently from each stratum. The sample sizes within each stratum are denoted by n_1, n_2, \dots, n_H , and the total PSU sample size is $n = \sum_{h=1}^H n_h$. At the second stage, a sample of m_{hi} secondary sampling units (SSU's) are selected from PSU i of size M_{hi} within stratum h by SRSWOR. At the third stage, a sample of l_{hij} ultimate sampling units (USU's) are selected from SSU ij of size L_{hij} within stratum h by SRSWOR. A vector of measurements of some unit characteristics is represented as $\mathbf{y}_{hijk} = (y_{1hijk}, y_{2hijk}, y_{3hijk})^T$, where the subscripts $hijk$ refer to the stratum label, PSU label, SSU label and USU label, respectively. The population parameter of interest $\theta = \theta(S)$, where $S = \{\mathbf{y}_{hijk} : h = 1, 2, \dots, H; i = 1, 2, \dots, N_h;$

1. F. Funaoka, Professor, Faculty of Economics, Shinshu University, 3-1-1 Asahi, Matsumoto, Nagano, 390-8621, Japan; H. Saigo, Professor, School of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda Shinjuku, Tokyo, 169-8050, Japan; R.R. Sitter, Professor, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, B.C., V5A 1S6, Canada; T. Toida, Associate Professor, Faculty of Social and Information Studies, Gunma University, 2-4 Aramakicho, Maebashi, Gunma 371-8510, Japan.

$j = 1, \dots, M_{hi}; k = 1, \dots, L_{hij}\}$, is usually estimated by $\hat{\theta} = \hat{\theta}(s)$, where $s = \{y_{hijk} : h = 1, \dots, H; i = 1, 2, \dots, n_h; j = 1, \dots, m_{hi}; k = 1, \dots, l_{hij}\}$. The population total vector is denoted $\mathbf{Y} = (Y_1, \dots, Y_r)'$. In this case, its unbiased estimate is

$$\hat{\mathbf{Y}} = \sum_{h=1}^H \hat{\mathbf{Y}}_h = \sum_{h=1}^H (N_h / n_h) \sum_{i=1}^{n_h} \hat{\mathbf{Y}}_{hi},$$

where $\hat{\mathbf{Y}}_{hi} = (M_{hi} / m_{hi}) \sum_{j=1}^{m_{hi}} \hat{\mathbf{Y}}_{hij}$ and $\hat{\mathbf{Y}}_{hij} = (L_{hij} / l_{hij}) \sum_{k=1}^{l_{hij}} y_{hijk}$. This may be written as $\hat{\mathbf{Y}} = \sum_{hijk} w_{hijk} y_{hijk}$, where $w_{hij} = (N_h / n_h) (M_{hi} / m_{hi}) (L_{hij} / l_{hij})$.

For $\tau = 1$, an unbiased estimate of $\text{Var}(\hat{\mathbf{Y}})$ is $v(\hat{\mathbf{Y}}) = \sum_{h=1}^H v(\hat{\mathbf{Y}}_h)$, where

$$v(\hat{\mathbf{Y}}_h) = \frac{N_h^2 (1 - f_{1h}) s_h^2}{n_h} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}^2 (1 - f_{2hi}) s_{hi}^2}{m_{hi}} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} \frac{L_{hij}^2 (1 - f_{3hij}) s_{hij}^2}{l_{hij}}$$

with $\bar{\mathbf{Y}}_h = n_h^{-1} \sum_i \hat{\mathbf{Y}}_{hi}$, $\bar{\mathbf{Y}}_{hi} = m_{hi}^{-1} \sum_j \hat{\mathbf{Y}}_{hij}$, $\bar{y}_{hijk} = l_{hij}^{-1} \sum_k y_{hijk}$, $f_{1h} = n_h / N_h$, $f_{2hi} = m_{hi} / M_{hi}$, $f_{3hij} = l_{hij} / L_{hij}$, $s_h^2 = \sum_i (\hat{\mathbf{Y}}_{hi} - \bar{\mathbf{Y}}_h)^2 / (n_h - 1)$, $s_{hi}^2 = \sum_j (\hat{\mathbf{Y}}_{hij} - \bar{\mathbf{Y}}_{hi})^2 / (m_{hi} - 1)$, and $s_{hij}^2 = \sum_k (y_{hijk} - \bar{y}_{hijk})^2 / (l_{hij} - 1)$ (Särndal, Swensson and Wretman 1992, pages 148–149).

3. Proposed Bernoulli Bootstrap

To handle the multi-stage aspect of the sampling within stratum, we propose a multi-stage bootstrap. To simplify ideas, we first introduce a simple version that has some limitations in applicability. We will then subsequently describe a more general form that avoids these difficulties.

A Short Cut BBE

Step I. For each sample PSU, hi , within stratum h , $h = 1, \dots, H$, we: (a) keep it in the bootstrap sample with probability

$$p_h = \sqrt{1 - \frac{(1 - f_{1h})}{(1 - n_h^{-1})}}; \quad (3.1)$$

or (b) replace it with one selected randomly from the n_h PSU's. If (a) is the case, go to Step II.

Step II. For each SSU hij in PSU hi of stratum h kept at Step I, we: (c) keep it in a bootstrap sample with probability

$$q_{hi} = \sqrt{1 - \frac{f_{1h}}{p_h^{-1}} \frac{(1 - f_{2hi})}{(1 - m_{hi}^{-1})}}; \quad (3.2)$$

or (d) replace it with one selected randomly from the m_{hi} SSU's in PSU hi of stratum h . If (c) is the case, go to Step III.

Step III. For each USU $hijk$ in SSU hij in PSU hi of stratum h , we: (e) keep it in the bootstrap sample with probability

$$r_{hij} = \sqrt{1 - \frac{f_{1h}}{p_h^{-1}} \frac{f_{2hi}}{q_{hi}^{-1}} \frac{(1 - f_{3hij})}{(1 - l_{hij}^{-1})}}; \quad (3.3)$$

or (f) replace it with one randomly selected from the l_{hij} USU's in SSU hij in PSU hi of stratum h .

If we let K_{hij}^* denote the number of times unit $hijk$ appears in the bootstrap resample, then the bootstrap estimate of the total is $\hat{\mathbf{Y}}^* = \sum_{hijk} w_{hij}^* y_{hijk}$, where $w_{hij}^* = K_{hij}^* w_{hij}$, and the bootstrap estimate of $V(\hat{\theta})$ is $v_B(\hat{\theta}) = V_*(\hat{\theta}^*)$, where $\hat{\theta}^* = \theta(\hat{\mathbf{Y}}^*)$ and V_* represents the variance under the resampling procedure. Typically, the bootstrap estimate of variance is obtained by Monte Carlo simulation. That is, repeat Steps I–III a large number of times, B , to get $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ and use

$$v_B(\hat{\theta}) = \sum_{b=1}^B (\hat{\theta}_b^* - \bar{\theta}_{(\cdot)}^*)^2 / B,$$

where $\bar{\theta}_{(\cdot)}^* = \sum_{b=1}^B \hat{\theta}_b^* / B$. In most cases one can replace $\bar{\theta}_{(\cdot)}^*$ by $\hat{\theta}$. This allows the survey methodologist to create a set of replicate weights w_{hij}^* for each bootstrap resample and release these with the data released to the public.

Obviously, the short cut BBE is feasible only when $p_h, q_{hi}, r_{hij} \in [0, 1] \forall h, i, j$. For instance, it is necessary that $f_{1h} \geq n_h^{-1}$. To handle arbitrary $n_h, m_{hi}, l_{hij} \geq 2$, we may modify each step and change p_h, q_{hi}, r_{hij} accordingly:

A General BBE

Step I'. Choose $(n_h - 1)$ PSU's by SRS with replacement from n_h PSU's in the sample, $h = 1, \dots, H$. Denote the candidate set by $\{\text{PSU}_{hi} : i = 1, 2, \dots, n_h - 1\}$. For each PSU i in the sample in stratum h , we: (a) keep it in the bootstrap sample with probability

$$p_h = 1 - \frac{1}{2} \frac{(1 - f_{1h})}{(1 - n_h^{-1})}; \quad (3.4)$$

or (b) replace it with one selected randomly from $\{\text{PSU}_{hi} : i = 1, 2, \dots, n_h - 1\}$. If (a) is the case, go to Step II'.

Step II'. For hi kept at Step I', choose $(m_{hi} - 1)$ SSU's by SRS with replacement from m_{hi} SSU's in PSU hi . Denote the candidate set by $\{\text{SSU}_{hij} : j = 1, 2, \dots, m_{hi} - 1\}$. For each SSU

hij in PSU hi kept at Step I', we: (c) keep it in the bootstrap sample with probability

$$q_{hi} = 1 - \frac{1}{2} \frac{f_{1h}}{p_h^{-1}} \frac{(1 - f_{2hi})}{(1 - m_{hi}^{-1})}; \quad (3.5)$$

or (d) replace it with one selected randomly from $\{\text{SSU}_{hij} : j = 1, 2, \dots, m_{hi} - 1\}$. If (c) is the case, go to Step III'.

Step III'. For hij kept at Step II', choose $l_{hij} - 1$ USU's by SRS with replacement from l_{hij} USU's in SSU hij in PSU hi . Denote the candidate set by $\{\text{USU}_{hijk} : k = 1, 2, \dots, l_{hij} - 1\}$. For each USU $hijk$ in SSU hij in PSU hi , we: (e) keep in the bootstrap sample with probability

$$r_{hij} = 1 - \frac{1}{2} \frac{f_{1h}}{p_h^{-1}} \frac{f_{2hi}}{q_{hi}^{-1}} \frac{(1 - f_{3hij})}{(1 - l_{hij}^{-1})}; \quad (3.6)$$

or (f) replace it with one randomly selected from $\{\text{USU}_{hijk} : k = 1, 2, \dots, l_{hij} - 1\}$.

It is readily seen that $p_h, q_{hi}, r_{hij} \in [0, 1] \forall n_h, m_{hi}, l_{hij} \geq 2$.

The reason for randomly selecting a candidate set in the general BBE can be explained as follows. To fix the idea, consider single-stratum one-stage SRSWOR. Let \bar{y}^* be a bootstrap sample mean under the short cut BBE with some arbitrary $p \in [0, 1]$. Then, it can be shown that $V_*(\bar{y}^*) = n^{-1}(1 - n^{-1})s^2(1 - p^2)$, where $s^2 = \sum_i (y_i - \bar{y})^2 / (n - 1)$. Note that $V_*(\bar{y}^*)$ is monotone decreasing with respect to p in $[0, 1]$. So, $\min_{p \in [0, 1]} V_*(\bar{y}^*) = 0$ and $\max_{p \in [0, 1]} V_*(\bar{y}^*) = n^{-1}(1 - n^{-1})s^2$. If $f_1 < n^{-1}$, then $\max_p V_*(\bar{y}^*) < v(\bar{y})$. The key idea of the general BBE is that we can make $\max_p V_*(\bar{y}^*)$ greater than $v(\bar{y})$ by putting extra variation into unit replacement through randomly selecting a candidate set.

It can be shown that both the short cut BBE and the general BBE provide consistent variance estimation for smooth functions of estimated population totals. Moreover, under appropriate regularity conditions for the population distribution function, they also provide consistent variance estimation for sample quantiles. In addition, both BBE methods use resample sizes equal to the original sample sizes. This can be a desirable property when we deal with imputed survey data (see Saigo, Shao and Sitter 2001).

It is not difficult to extend the BBE approach to designs with more than three stages. For example, for a four stage stratified design, a USU at the fourth stage within stratum h is kept with probability

$$\sqrt{1 - p_h^{-1} f_{1h} q_{hi}^{-1} f_{2hi} r_{hij}^{-1} f_{3hij} (1 - g_{hijk}^{-1})^{-1} (1 - f_{4hijk})}$$

or replaced in the short cut BBE, where g_{hijk} is the fourth stage sample size and f_{4hijk} is the fourth stage sampling fraction. Further extensions are analogous.

The general BBE randomizes a candidate set in order to merely fix infeasibility of the short cut BBE. This idea has similarities to the approximately Bayesian bootstrap of Rubin and Schenker (1986).

A disadvantage of the general BBE versus the short cut BBE is that the former requires, on the average, $\sum_h \{(n_h - 1) + p_h \sum_i (m_{hi} - 1) + p_h \sum_i q_{hi} \sum_j (l_{hij} - 1)\}$ more random number generations than the latter, where p_h, q_{hi} , and r_{hij} are given by (3.4), (3.5), and (3.6), respectively. This may be time-consuming when the sample sizes and/or the number of strata are large. To reduce random number generations in the general BBE, one can create a candidate set by randomly deleting one unit from the original sample and use

$$p_h = (n_h + 1/2) - \sqrt{(n_h + 1/2)^2 - n_h(1 + f_{1h})}, \quad (3.7)$$

$$q_{hi} = (m_{hi} + 1/2) - \sqrt{(m_{hi} + 1/2)^2 - f_{1h} p_h^{-1} m_{hi}(1 + f_{2hi})}, \quad (3.8)$$

$$r_{hij} = (l_{hij} + 1/2) - \sqrt{(l_{hij} + 1/2)^2 - f_{1h} p_h^{-1} f_{2hi} q_{hi}^{-1} l_{hij}(1 + f_{3hij})}, \quad (3.9)$$

instead. It can be shown that $p_h, q_{hi}, r_{hij} \in [0, 1]$. The proof for this modified version of the general BBE is similar.

4. A Simulation Study

In this section, we perform limited simulations to examine the BBE for ratio estimation and quantile estimation. For simplicity, we consider two-stage SRSWOR and restrict to a single stratum.

4.1 General Description of Simulation

A single-stratum finite population is generated by the following procedure and fixed over all simulation runs to observe design-based properties of the BBE. First, the average of the auxiliary variables in cluster i is generated by $\mu_i \sim N(\mu, \sigma^2)$ for $i = 1, 2, \dots, N$. Then, the auxiliary variable x_{ik} of unit k in cluster i is generated by

$$x_{ik} = \mu_i + \varepsilon_{ik} \quad (k = 1, 2, \dots, M_i; i = 1, 2, \dots, N), \quad (4.1)$$

where $\varepsilon_{ik} \sim N(0, (1 - \rho)\sigma^2 / \rho)$. The target variable y_{ik} of unit k in cluster i is obtained by

$$y_{ik} = a + b x_{ik} + e_{ik} \quad (k = 1, 2, \dots, M_i; i = 1, 2, \dots, N), \quad (4.2)$$

where $e_{ik} \sim N(0, \sigma^2/4)$. The parameter values used are $\mu = 100, \sigma = 10, \rho = 0.1(0.3), a = 0$, and $b = 1$, and two-stage SRSWOR is used throughout the simulation study.

4.2 Ratio Estimation

Let $N = 50$, $n = 15$, $M_i = 20$ and $m_i = 3$, for $i = 1, \dots, n$. Consider the ratio estimator of the population total, Y ,

$$\hat{Y}_R = \hat{R} X,$$

where $X = \sum_{i=1}^N \sum_{k=1}^{M_i} x_{ik}$ is the population total of the x 's $\hat{R} = \hat{Y} / \hat{X}$, $\hat{Y} = \sum_{h=1}^H \hat{Y}_h = \sum_{h=1}^H (N_h / n_h) \sum_{i=1}^{n_h} \hat{Y}_{hi}$, $\hat{X} = \sum_{h=1}^H \hat{X}_h = \sum_{h=1}^H (N_h / n_h) \sum_{i=1}^{n_h} \hat{X}_{hi}$, $\hat{Y}_{hi} = (M_{hi} / m_{hi}) \sum_{k=1}^{m_{hi}} \hat{Y}_{hik}$ and $\hat{X}_{hi} = (M_{hi} / m_{hi}) \sum_{k=1}^{m_{hi}} \hat{X}_{hik}$.

For the purpose of comparison, we consider a number of alternate variance estimators that are available in this simple context:

- 1) The conventional variance estimator is denoted

$$v_0(\hat{Y}_R) = N^2 \frac{1 - f_1}{n} \frac{\sum_i (\hat{Y}_i - \hat{R} \hat{X}_i)^2}{n - 1} + \frac{N}{n} \sum_i \frac{M_i^2 (1 - f_{2i}) s_{d'2i}^2}{m_i}, \quad (4.3)$$

where $f_1 = n / N$, $f_{2i} = m_i / M_i$ and

$$s_{d'2i}^2 = \sum_j (y_{ij} - \hat{R} x_{ij})^2 / (m_i - 1).$$

- 2) The delete 1 PSU at a time jackknife corrected for the first-stage sampling fraction is sometimes used, even though it is not entirely correct,

$$v_{cj}(\hat{Y}_R) = (1 - f_1) \frac{n - 1}{n} \sum_i (\hat{Y}_{R(i)} - \hat{Y}_{R(-)})^2, \quad (4.4)$$

where $\hat{Y}_{R(i)}$ is the estimator recalculated with the i^{th} PSU removed and $\hat{Y}_{R(-)} = \sum_i \hat{Y}_{R(i)} / n$.

- 3) An externally weighted jackknife (see Folsom, Bayless and Shah 1971) can be derived that corrects for both stages of sampling as

$$v_{ewj}(\hat{Y}_R) = (1 - f_1) \frac{n - 1}{n} \sum_i (\hat{Y}_{R(i)} - \hat{Y}_{R(-)})^2 + f_1 \sum_i (1 - f_{2i}) \frac{m_i - 1}{m_i} \sum_j (\hat{Y}_{R(ij)} - \hat{Y}_{R(-)})^2, \quad (4.5)$$

where $\hat{Y}_{R(i)}$ is the i^{th} jackknife pseudo value by deleting PSU i , $\hat{Y}_{R(ij)}$ is the ij^{th} jackknife pseudo value by deleting unit j in PSU i , $\hat{Y}_{R(-)} = \sum_i \hat{Y}_{R(i)} / n$, and $\hat{Y}_{R(-)} = \sum_j \hat{Y}_{R(ij)} / m_i$.

- 4) A model-assisted variance estimator is also available (see Särndal, Swensson and Wretman (1992), equation (8.10.6)),

$$v_{ma}(\hat{Y}_R) = (X / \hat{X})^2 v_0(\hat{Y}_R). \quad (4.6)$$

We use $B = 100$ bootstrap resamples in each of $S = 1,000$ simulation runs. The true MSE's are approximated by 10,000 simulation runs and we use Monte

Carlo estimates of percent relative bias and coefficient of variation of the various variance estimators as measures of their relative performance, as well as, empirical coverage probabilities of 90% confidence intervals.

We see in Table 1 that v_{BBE} , v_0 , v_{ewj} and v_{ma} perform comparably and well, except that the CV of the resampling methods are a bit higher than the non-resampling methods, as is typical. The delete 1 PSU at a time jackknife performs poorly.

To investigate the conditional properties, we ordered the 1,000 simulation runs on X / \hat{X} and divided the runs into 20 equally sized groups. For each group the average of each variance estimator is calculated. Figure 1 plots these grouped averages for each variance estimator (excluding v_{cj} since it has large negative bias) versus the grouped average X / \hat{X} , for $\rho = 0.3$. The true MSE is included in the plot, as well. This is a similar plot to that used by Royall and Cumberland (1981a, 1981b). One can see that v_{BBE} tracks the true MSE much like v_{ewj} and v_{ma} , whereas v_0 does not. Thus, the BBE seems to have a desirable conditional property.

Table 1
Comparison of Variance Estimators for \hat{Y}_R

ρ		% Bias	CV	Coverage (90%)
0.1	v_0	-1.70	0.28	89.2
	v_{BBE}	-0.62	0.33	88.9
	v_{ewj}	-0.33	0.30	89.4
	v_{cj}	-26.55	0.39	80.5
	v_{ma}	-0.39	0.30	89.4
0.3	v_0	-0.67	0.28	86.6
	v_{BBE}	-1.63	0.33	86.5
	v_{ewj}	-0.74	0.29	86.5
	v_{cj}	-26.85	0.39	80.2
	v_{ma}	-0.87	0.29	86.4

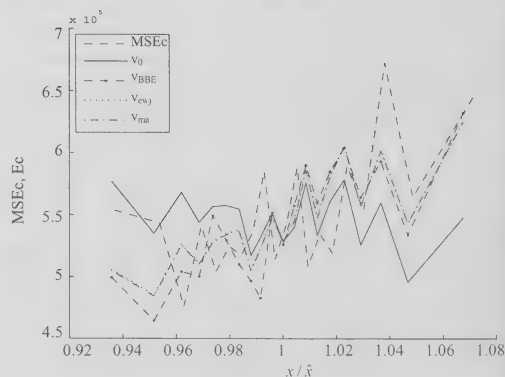


Figure 1. MSEc and Ec(v) for the ratio estimation.

4.3 Quantile Estimation

For quantile estimation, we set $N=100, n=30, M_i=100$ and $m_i=10$, for $i=1, \dots, n$. We use $B=500$ bootstrap resamples in each of $S=5,000$ simulation runs. The true MSE's are approximated by 50,000 simulation runs. Only the results for v_{BBE} and v_{ewj} when $\rho=0.1$ are summarized in Table 2 because those when $\rho=0.3$ are similar. We see that the BBE method performs quite well, with a slight upward bias, while the externally weighted jackknife method has serious bias because of its inconsistency in variance estimation for quantiles.

Table 2
Performance of v_{BBE} and v_{ewj} for the 0.10, 0.25, 0.50, 0.75 and 0.90 quantiles

Quantile	v_{BBE}			v_{ewj}		
	%Bias	CV	Coverage (90%)	%Bias	CV	Coverage (90%)
0.10	8.40	0.51	87.7	51.87	1.93	81.3
0.25	6.21	0.42	88.2	21.19	1.28	83.3
0.50	2.53	0.37	87.4	14.27	1.00	83.0
0.75	6.23	0.42	87.8	28.07	1.33	83.4
0.90	6.32	0.50	88.0	54.47	2.05	80.3

5. Application to the 1997 National Survey of Prices in Japan

The objective of the NSP is to analyze price formations of major consumers' goods, such as food, clothes and home appliances. To this end, quantile estimation plays a central role, and many quantile estimates based on several post-stratifications are included in the NSP reports.

The stratified multistage sampling used in NSP 1997 is summarized as follows:

Stratification. Municipalities form the PSU's and are stratified into 537 strata, first according to prefectures and economic sphere that each municipality forms and then further by their population sizes.

First Stage Sampling. These PSU's are selected via SRSWOR independently within each stratum. An overview of the first-stage sampling fractions is given in Table 3.

Second Stage Sampling. In a selected municipality, all the large scale outlets are enumerated. In other words, single stage cluster sampling is employed for large scale outlets. For small scale outlets, on the other hand, a sampled municipality is divided into survey areas (SSU's) each

consisting of about 100 outlets. Systematic sampling is used to sample survey areas. The sampling fractions at the second stage are between 0.1 and 1.0.

Third Stage Sampling. In each selected survey area, 40 outlets (USU's) are chosen by ordered systematic sampling with respect to the types of outlets and the annual sales reported in the 1994 Census of Commerce.

Strictly speaking, there is no valid variance formula for the NSP data because it contains systematic sampling. For estimating variance, however, it is assumed that systematic sampling can be approximated by SRSWOR. Even under this simplified condition, there is no closed variance formula for sample quantiles. In fact, no variance estimates are associated with estimated price quantiles in the NSP report, while the average prices are reported with their variance estimates.

In this section, we apply the short cut BBE to the NSP data, assuming that systematic sampling can be approximated by SRSWOR. Some strata have only one PSU. In addition, $f_{1h} < n_h^{-1}$ in some strata. Such strata are grouped into adjacent strata so that p_h given by (3.1) is in $[0, 1]$. After grouping, there are more than 280 strata. The effect of reforming strata is assumed to be negligible.

Table 3
The First Stage Sampling Fractions in NSP 1997

Area Category	Population Size	# of PSU's	Sampling Fraction	Sample Size
Cities	$\geq 100,000$	221	1/1	221
Cities	50,000 – 99,999	220	2/3	179
Cities	$< 50,000$	224	1/3	80
Towns and villages	$\geq 40,000$	32	1/5	4
Towns and villages	$< 40,000$	2,536	1/15	187

After reforming strata, the short cut BBE is employed in those strata composed by cities. On the other hand, the with-replacement bootstrap (Shao and Tu 1995, page 247) using resample size $(n_h - 1)$ is used in those composed of towns and villages, where the first stage sampling fractions are small. The quantile estimates and their standard errors for selected commodities in small-sized outlets are shown in Table 4. Note that the prices of a given commodity are discrete. However, we apply the bootstrap as if prices of commodities are continuous. This approximation should be acceptable for many commodities, but not for very inexpensive ones, since in such a case, a large percentage of observations concentrate on a specific price and the estimated standard error can be 0.

Table 4
Sample Quantiles (Standard Errors) of Selected Commodities for Small Outlets in NSP

Commodity	p	0.10	0.25	0.5	0.75	0.90
Rice (5kg) ^a (10 yens)	Sample quantile	239.4	255.2	278.3	299.1	315.0
	(standard error)	(0.24)	(0.53)	(0.21)	(0.02)	(0.61)
Instant Coffee (1 bottle) ^b (yen)	Sample quantile	714	788	859	893	914
	(standard error)	(0.13)	(0.40)	(0.00)	(2.68)	(1.43)
Beer (24 cans) ^c (10 yens)	Sample quantile	467.3	500.0	536.8	549.4	549.4
	(standard error)	(1.01)	(0.64)	(0.82)	(0.00)	(0.00)
PC ^d (1,000 yens)	Sample quantile	248.8	260.4	299.3	346.5	375.9
	(standard error)	(2.03)	(0.35)	(3.25)	(7.17)	(1.48)

The specified brands ^aKoshihikari; ^bNescafe Gold Blend, 100g; ^cSapporo (Nama) Black Label, 350ml;

^dNEC PC9821 NW133/D14.

6. Conclusions

The bootstrap is useful for estimating variances in complex surveys, particularly when quantile estimation is important. We have proposed two Bernoulli-type bootstrap methods that can easily handle multi-stage stratified SRSWOR designs where sampling fractions are large: the short cut BBE and the general BBE. In both methods, a sampling unit at a given stage is either kept or replaced with preassigned probabilities to construct a bootstrap sample. The general BBE has an advantage in that it can handle any combination of sample sizes ≥ 2 although it requires more random number generations than the short cut BBE. As an illustration, we applied the short cut BBE to Japanese 1997 National Survey of Prices data.

Acknowledgements

The second author was supported by the Japan Statistical Association. The third author was supported by a grant from the Natural Science and Engineering Research Council of Canada. The authors thank the Statistics Bureau, Ministry of Public Management, Home Affairs, Posts and Telecommunications, and the Ministry of Economy, Trade, and Industry, Japan, for providing the 1997 NSP data.

References

- Bickel, P.J., and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.
- Booth, J.G., Butler, R.W. and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282-1289.
- Folsom, R.E., Bayless, D.L. and Shah, B.V. (1971). Jackknifing for variance components in complex sample survey designs. *Proceedings of the Social Statistics Section*, American Statistical Association, 36-39.
- Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 181-184.
- Kovar, J.G., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, Supplement, 25-45.
- McCarthy, P.J., and Snowden, C.B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics*, Serie 2, 95, Public Health Service Publication, 85-1369, Washington, DC: U.S. Government Printing Office.
- Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.
- Royall, R.M., and Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.
- Royall, R.M., and Cumberland, W.G. (1981b). The finite population linear regression estimator: An empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- Rubin, D.B., and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- Saigo, H., Shao, J. and Sitter, R.R. (2001). A repeated half-sample bootstrap and balanced repeated replication for randomly imputed data. *Survey Methodology*, 27, 189-196.
- Särndal, C.-E., Swenson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Shao, J., and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.
- Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.
- Sitter, R.R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.
- Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20, 135-154.

Geometric Versus Optimization Approach to Stratification: A Comparison of Efficiency

Marcin Kozak and Med Ram Verma¹

Abstract

In this paper, the geometric, optimization-based, and Lavallée and Hidirolou (LH) approaches to stratification are compared. The geometric stratification method is an approximation, whereas the other two approaches, which employ numerical methods to perform stratification, may be seen as optimal stratification methods. The algorithm of the geometric stratification is very simple compared to the two other approaches, but it does not take into account the construction of a take-all stratum, which is usually constructed when a positively skewed population is stratified. In the optimization-based stratification, one may consider any form of optimization function and its constraints. In a comparative numerical study based on five positively skewed artificial populations, the optimization approach was more efficient in each of the cases studied compared to the geometric stratification. In addition, the geometric and optimization approaches are compared with the LH algorithm. In this comparison, the geometric stratification approach was found to be less efficient than the LH algorithm, whereas efficiency of the optimization approach was similar to the efficiency of the LH algorithm. Nevertheless, strata boundaries evaluated via the geometric stratification may be seen as efficient starting points for the optimization approach.

Key Words: Optimum stratification; Geometric Stratification; Numerical Optimization; Lavallée-Hidirolou algorithm.

1. Introduction

Gunning and Horgan (2004) proposed a stratification algorithm based on a geometric progression. For the sake of simplicity, we will call this technique the “geometric approach to stratification,” “geometric stratification,” or just “geometric approach.” The geometric stratification aims to equalize values of the coefficient of variation of a stratification variable within strata, based on the assumption that the variable is uniformly distributed within each stratum. Gunning and Horgan (2004) showed that their algorithm is much easier to implement and more efficient than the classical cumulative root frequency method (Dalenius and Hodges 1959) as well as the Lavallée and Hidirolou (LH) algorithm (Lavallée and Hidirolou 1988). Horgan (2006) compared the geometric stratification with the Dalenius and Hodges’ (1959), Ekman’s (1959), and Lavallée and Hidirolou (1988) procedures; again, in their study the geometric stratification occurred to be the most efficient among the procedures compared. Gunning, Horgan and Yancey (2004) applied this method to stratify accounting populations.

Like the cumulative square root frequency method, the geometric approach is an approximate stratification technique, and hence the stratification points it provides may be quite far from optimum stratification points. On the other hand, there exist approaches, especially for univariate stratification, that lead to near-optimum stratification points.

These approaches are based on the use of self-implemented algorithms or numerical optimization methods to provide strata boundaries (e.g., Lavallée and Hidirolou 1988; Lednicki and Wieczorkowski 2003; Kozak 2004). Such methods, however, usually require initial strata boundaries to start an optimization process; approximate stratification methods can be employed to find such initial points. Of course, initial strata boundaries should be of high quality, as their low quality may cause the optimization to provide a local minimum (Rivest 2002).

Many surveys deal with positively skewed study variables. If this is the case, it is important to take into account this attribute when stratifying a population. Many researchers have attempted to create stratification methods that would construct a so-called “take-all” stratum (e.g., Glasser 1962; Hidirolou 1986), from which all the elements are selected in the sample with probability 1. In stratified sampling, this is the best manner of dealing with positively skewed variables. Such methods are usually more efficient (certainly, only if a population is positively skewed) than stratification methods in which a take-all stratum is not constructed. A take-all stratum is not constructed in the geometric stratification (Gunning and Horgan 2004).

The aim of this paper is to compare the efficiency of the geometric stratification, proposed by Gunning and Horgan (2004), and two optimization approaches to stratification (Lavallée and Hidirolou 1988; Lednicki and

1. Marcin Kozak, Department of Biometry, Warsaw Agricultural University, Nowoursynowska 159, 02-776 Warsaw, Poland. E-mail: marcin.kozak@omega.sggw.waw.pl; Med Ram Verma, Division of Agricultural Economics & Statistics, ICAR Research Complex for N.E.H. Region, Umroi Road, Umiam (Barapani) Meghalaya, India, Pin 793 103. E-mail: mrverma19@yahoo.co.in.

Wieczorkowski 2003; Kozak 2004), which is based on the use of numerical optimization methods.

2. Stratification Approaches Compared

Suppose we aim to stratify an N -element positively skewed population, U , based on an N -vector $\mathbf{x} = (x_1, \dots, x_N)^T$ of values, known at the outset (*i.e.*, prior to the study), of a stratification variable X .

In this paper, we consider two stratification problems. In the first problem, L strata are to be constructed subject to a given sample size n . Suppose we are looking for an $(L+1)$ -vector of strata boundaries $\mathbf{k} = (k_0, \dots, k_L)^T$ ($k_0 < k_1 < \dots < k_L$, k_0 being the minimum and k_L the maximum value of X) that minimizes the variance of an estimator of the population mean of X under stratified sampling with simple random sampling without replacement within strata (*STSI*) sampling combined with a take-all stratum approach. (Note that we treat the stratification variable as identical to the corresponding survey variable.) The variance of \bar{x}_{st} is given by

$$V(\bar{x}_{st}) = \sum_{h=1}^{L-1} \left(\frac{N_h}{N} \right)^2 \left(1 - \frac{n_h}{N_h} \right) \frac{S_h^2}{n_h},$$

$$\bar{x}_{st} = \sum_{h=1}^L \frac{N_h}{N} \bar{x}_h, \bar{x}_h = \frac{1}{n_h} \sum_{k=1}^{n_h} x_{kh} \quad (h=1, \dots, L), \quad (1)$$

where n_h is the sample size from the h^{th} stratum, N_h is the size of the h^{th} stratum, S_h^2 is the population variance of X restricted to the h^{th} stratum, \bar{x}_{st} is the estimator of the population mean of X under *STSI* sampling, \bar{x}_h is the estimator of the population mean of X in the h^{th} stratum under simple random sampling without replacement (*SI*) sampling, and x_{kh} is the value of X for the k^{th} sample element of the h^{th} stratum and $h=1, \dots, L$.

The optimum sample allocation, which is in our problem obtained by minimizing the variance (1) subject to a given sample size n , is given by the following Neyman-optimum formula adjusted to a take-all stratum approach (Lednicki and Wieczorkowski 2003):

$$n_h = (n - N_L) \frac{N_h S_h}{\sum_{h=1}^{L-1} N_h S_h}, \quad h=1, \dots, L-1. \quad (2)$$

The geometric approach to stratification aims to equalize values of the coefficient of variation of X within the L strata. It simply consists of applying the following formula based on a geometric progression (Gunning and Horgan 2004)

$$k_h = ar^h, \quad h=0, \dots, L, \quad (3)$$

where $a = \min(X)$, $k_L = \max(X)$, and $r = (k_L/k_0)^{1/L}$. The formula (3) is based on the assumption that X is uniformly distributed within each stratum.

The optimization approach applied to this particular stratification problem is based on the numerical optimization of the following problem: Minimize

$$f(\mathbf{k}) = V(\bar{x}_{st}), \quad (4)$$

where $V(\bar{x}_{st})$ is the variance (1) under the optimum allocation (2), subject to constraints

$$N_h \geq 2 \text{ and } 2 \leq n_h \leq N_h \text{ for } h=1, \dots, L-1, \quad (5)$$

and

$$\sum_{h=1}^{L-1} n_h = n - N_L. \quad (6)$$

Sometimes, when one wants to obtain more or less equal levels of precision of estimation in each stratum, a power allocation may be applied (Bankier 1988; Rivest 2002; Lednicki and Wieczorkowski 2003):

$$n_h = \frac{(n - N_L)(N_h \bar{x}_h)^p}{\sum_{h=1}^{L-1} (N_h \bar{x}_h)^p}, \quad p \in (0,1]; \quad h=1, \dots, L-1. \quad (7)$$

The optimization approach is more difficult to apply than the geometric stratification approach due in large part to the fact that the algorithm for the geometric approach is significantly more simplistic than for the optimization approach. An optimization method has to be chosen from among various available methods. Lednicki and Wieczorkowski (2003) used the simplex method of Nelder and Mead (1965); however, more efficient methods, which often require self-implemented algorithms (*e.g.*, Kozak 2004), can be applied, too.

Note that the geometric stratification does not take into account the formulae for the variance (1), the sample allocation (2), and the constraints (5). It may happen that one of the constraints (5) is not fulfilled. For these reasons, the geometric stratification is an approximate stratification procedure.

In this study, the algorithm proposed by Kozak (2004) was applied to stratify several populations. It is a random search algorithm adjusted to the problem of stratification. It is a simple algorithm; in each step, a stratum boundary is randomly selected and randomly changed. If the new set of strata boundaries is better than the previous one, the new one replaces the previous one. In the Appendix, the algorithm based on the paper by Kozak (2004) is given in detail.

The second problem considered in the paper is construction of strata that minimize a sample size from a population with respect to a given level of precision of estimation (the precision of estimation being given by the variance of an estimator of the population mean or total). The Lavallée-Hidiroglou (LH) algorithm (Lavallée and Hidiroglou 1988) can be seen as a particular optimization method to solve this particular stratification problem; it does not, however, work in other problems, e.g., in the one considered earlier. For details of the algorithm, see the paper by Lavallée and Hidiroglou (1988). Besides the LH algorithm, the geometric stratification and random search method were applied to construct the strata.

The R language and environment (R Development Core Team 2005) was used to perform all the computation work in the present study.

3. Numerical Comparison of Efficiency of the Approaches in Stratification Under Fixed Sample Size

In this section, we compare two stratification approaches, the geometric stratification (geom) and optimization approach (optim), applied to a problem of searching for the strata boundaries that minimize the variance of the considered estimator with respect to a fixed sample size. In order to perform the comparison, five artificial populations of various sizes (from 2,000 to 10,000) were generated. Their summary statistics are presented in Table 1; the histograms of the stratification variables in the populations are given in Figure 1. In each case, the stratification variable was positively skewed (the skewness ranged between 1.40 for the 1st population to 5.02 for the 5th population). As it is usually the case in real populations, values of the stratification variables were integers. The sample size, n_i , from the i^{th} population was $n_i = f N_i$, where $f = 0.15$ is an assumed sample fraction and N_i is the size of the i^{th} population.

Table 1
Summary Statistics for Studied Artificial Populations

Population	Size	Range	Skewness	Mean	Variance
1	4,000	3–72	1.40	16.11	45.8
2	4,000	243–28,578	2.66	2,823.95	4.8×10^6
3	2,000	6–2,793	3.55	224.12	6.0×10^4
4	10,000	62–74,398	4.20	3,616.41	2.1×10^7
5	2,000	259–186,685	5.02	9,265.36	1.1×10^8

First, each population was stratified using the geometric stratification method into 4, 5, 6, and 7 strata. Then, the optimization approach was applied; as initial parameters in the optimization approach, the strata boundaries determined via the geometric stratification were used.

Like Gunning and Horgan (2004), to compare the efficiency of the two approaches, the relative efficiency was calculated via the formula:

$$\text{eff}_{\text{geom, optim}} = \frac{V_{\text{geom}}(\bar{x}_{\text{st}})}{V_{\text{optim}}(\bar{x}_{\text{st}})}, \tag{8}$$

where $V_{\text{geom}}(\bar{x}_{\text{st}})$ and $V_{\text{optim}}(\bar{x}_{\text{st}})$ are the variances (1) under the geometric and optimization approach, respectively. In addition, we calculated the coefficients of variation of the estimator of the population mean under both approaches:

$$\text{cv}_{\text{geom}} = \frac{\sqrt{V_{\text{geom}}(\bar{x}_{\text{st}})}}{\bar{x}_{\text{st}}}; \text{cv}_{\text{optim}} = \frac{\sqrt{V_{\text{optim}}(\bar{x}_{\text{st}})}}{\bar{x}_{\text{st}}}. \tag{9}$$

Table 2 contains the values of the relative efficiencies (8) and the coefficients of variation (9) for each combination studied (population \times number of strata).

Table 2
Coefficients of Variation of the Estimator of the Population Mean Under the Geometric Stratification (CV_{geom}) and Optimization Approach (CV_{optim}), and Efficiencies of the Geometric Stratification Relative to the Optimization Approach ($\text{eff}_{\text{geom, optim}}$)

Number of strata L	CV_{geom}	CV_{optim}	$\text{eff}_{\text{geom, optim}}$
Population 1			
4	0.0086	0.0056	1.53
5	0.0070	0.0042	1.66
6	0.0057	0.0034	1.66
7	0.0051	0.0029	1.75
Population 2			
4	0.0116	0.0084	1.37
5	0.0095	0.0065	1.47
6	0.0085	0.0051	1.66
7	0.0073	0.0042	1.72
Population 3			
4	0.0235	0.0133	1.76
5	0.0174	0.0100	1.74
6	0.0146	0.0081	1.80
7	0.0129	0.0067	1.91
Population 4			
4	0.0104	0.0063	1.64
5	0.0089	0.0047	1.88
6	0.0073	0.0038	1.93
7	0.0064	0.0032	2.00
Population 5			
4	0.0235	0.0134	1.76
5	0.0185	0.0100	1.86
6	0.0161	0.0080	2.00
7	0.0134	0.0074	1.82

In each case, the optimization approach was more efficient than the geometric stratification. The efficiency was smaller than 1.5 for only two combinations; in the rest of combinations, it ranged between 1.5 and 2. Usually, the more strata constructed the greater the gain in efficiency.

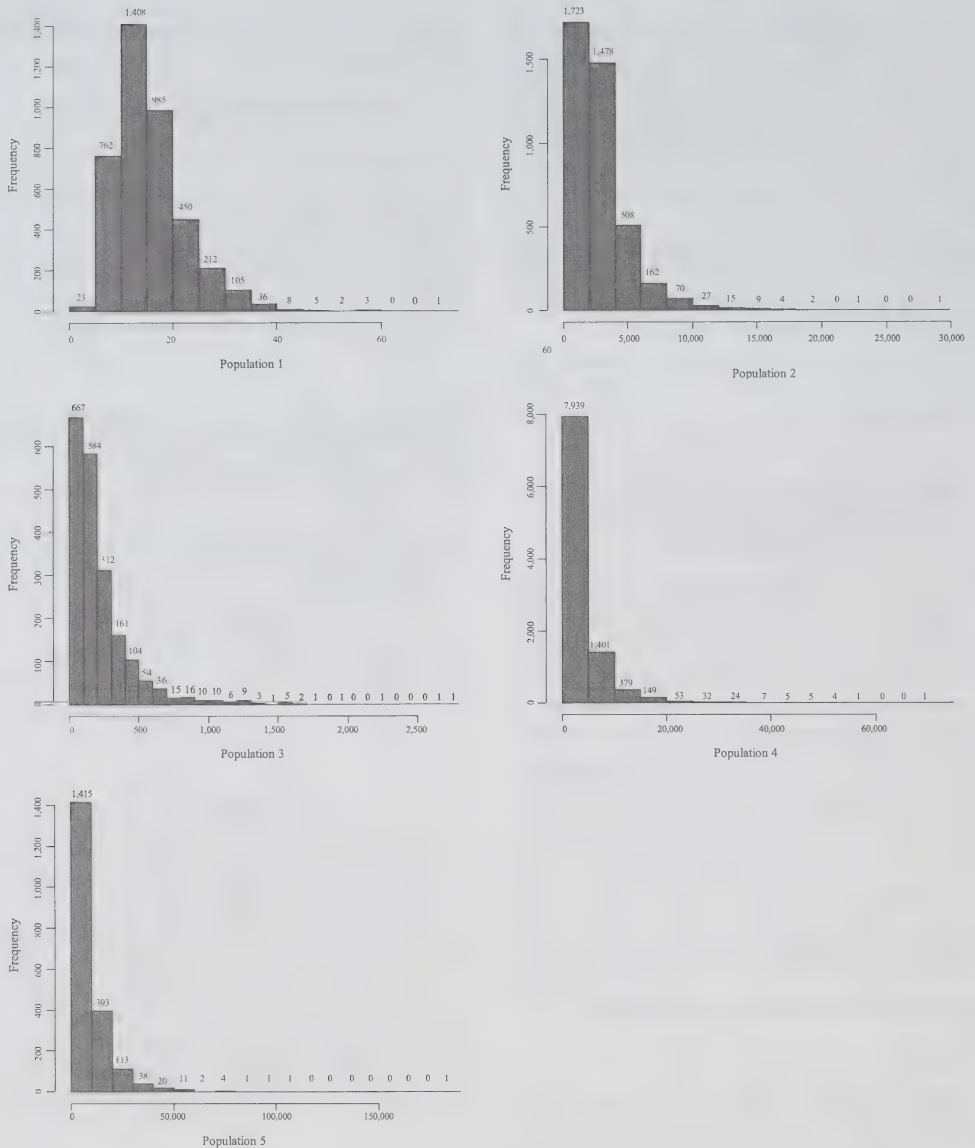


Figure 1. Histograms of stratification variable in studied artificial populations.

4. Numerical Comparison of Efficiency of the Stratification Approaches Under Fixed Level of Precision of Estimation

Gunning and Horgan (2004) and Horgan (2006) compared the geometric stratification with the Lavallée and Hidroglou (Lavallée and Hidroglou 1988) algorithm and

found that the former was usually more efficient. In this section, we compare the three stratification approaches: the geometric stratification, the LH algorithm, and the optimization approach via a random search method. In this study, the same five populations as in the previous section were used (see Table 1 and Figure 1).

The relative efficiencies of two approaches were evaluated as

$$\text{eff}_{i,j} = \frac{n_j(\text{cv})}{n_i(\text{cv})}, \tag{10}$$

where i and j are the indices of the stratification approaches ($i, j = \text{geom, optim, LH}$), and $n_i(\text{cv})$ and $n_j(\text{cv})$ are the minimum sample sizes required to obtain a desired level of precision (cv) under the i^{th} and j^{th} approaches, respectively.

Using the three approaches, each population was stratified into $L = 4, \dots, 7$ strata; the required level of precision was 0.01 in each case. Minimum sample sizes required for this level of precision and relative efficiencies (10) are given in Table 3.

Table 3

Minimum Sample Sizes Required to Obtain a Value Equal to 0.01 for the Coefficient of Variation of the Estimator of the Population Mean, Under the Geometric Stratification (n_{geom}), Optimization Approach (n_{optim}), and LH Algorithm (n_{LH}); and Efficiencies of the Geometric Stratification Relative to the Optimization Approach ($\text{eff}_{\text{geom, optim}}$), the Geometric Stratification Relative to the LH Algorithm ($\text{eff}_{\text{geom, LH}}$), and LH Algorithm Relative to the Optimization Approach ($\text{eff}_{\text{LH, optim}}$)

Number of strata L	n_{geom}	n_{optim}	n_{LH}	$\text{eff}_{\text{geom, optim}}$	$\text{eff}_{\text{geom, LH}}$	$\text{eff}_{\text{LH, optim}}$
Population 1						
4	805	496	496	1.63	1.63	1.00
5	613	344	344	1.78	1.78	1.00
6	460	252	252	1.83	1.83	1.00
7	357	192	192	1.86	1.86	1.00
Population 2						
4	483	248	259	1.94	1.86	1.04
5	329	154	163	2.14	2.02	1.06
6	224	113	117	1.98	1.92	1.03
7	180	83	83	2.17	2.17	1.00
Population 3						
4	782	410	411	1.91	1.90	1.00
5	601	303	304	1.98	1.98	1.00
6	495	242	241	2.04	2.05	1.00
7	422	195	195	2.11	2.16	1.00
Population 4						
4	839	409	409	2.05	2.05	1.00
5	650	301	301	2.15	2.15	1.00
6	552	240	242	2.30	2.28	1.01
7	— ¹	200	200	—	—	1.00
Population 5						
4	1,768	894	894	1.98	1.98	1.00
5	1,274	628	628	2.03	2.03	1.00
6	949	459	459	2.07	2.07	1.00
7	758	355	355	2.13	2.13	1.00

¹ There were numerical problems with obtaining stratum boundaries (sample sizes from some strata were bigger than the sizes of these strata).

From the results it follows that the optimization approach was more efficient than the geometric stratification; this outcome was obtained for each population and number of strata. The relative efficiency was always greater than 1.6. Moreover, an interesting conclusion follows from the comparison of the efficiency of the geometric and LH

stratifications. As already mentioned, Gunning and Horgan (2004) and Horgan (2006) found the geometric stratification more efficient than the LH algorithm. On the contrary, in our study, the LH algorithm was always more efficient than the geometric stratification. This situation occurred also for other generated populations of various sizes and skewness (results not included in this paper). Nevertheless, we do not state that the LH algorithm is always more efficient than the geometric stratification. It may happen that the geometric stratification will be better, as Gunning and Horgan (2004) and Horgan (2006) obtained in their studies.

From the comparison of the LH algorithm and the optimization approach it follows that both approaches provides stratification points leading to similar sample sizes. In some cases, the LH stratification was slightly better and in some other cases slightly worse than the optimization approach. Nevertheless, these differences do not mean that we could indicate either of these two approaches as more efficient. In fact, these two approaches have the same aim (in this particular stratification problem) and they just differ in the algorithm to achieve this aim. In summary, on the basis of our results we conclude that, in general, the LH stratification and optimization approach are more efficient than the geometric stratification.

5. Conclusions

The stratification technique based on a geometric progression proposed by Gunning and Horgan (2004) has a significant advantage; namely, its algorithm is very simple to implement compared to the cumulative square root of frequency method of Dalenius and Hodges (1959) and to other stratification methods. It is, however, an approximate stratification procedure, so the stratification points it provides may lead to poor precision of estimation (or a large sample size required to achieve a required level of precision). Furthermore, it is likely that some of the strata constructed will not fulfill the constraints (5); e.g., some strata may be empty (so they would not comprise any population element) or/and sample sizes from some strata may be smaller than two or greater than their population sizes.

In our study, the optimization approach (via the LH and random search algorithms) was more efficient than the geometric stratification for each population studied and number of strata constructed. Nevertheless, the strata boundaries provided by the geometric stratification can be seen as efficient initial parameters required in the optimization approach; they should not be considered, however, as the optimal or efficient strata boundaries. Furthermore, our results conclusively show that the geometric stratification is less efficient than the stratification

presented by Lavallée and Hidirolou (1988), which is the result opposite to the one obtained by Gunning and Horgan (2004) and Horgan (2006). This problem needs further studies on real skewed populations; investigations on artificial populations univocally show that the LH algorithm and the optimization approach are more efficient than the geometric stratification.

At first look, one could be surprised that the gain in efficiency after applying the LH and optimization approaches compared to the geometric stratification increases after increasing the number of strata. This can be easily explained. The aim of the geometric stratification is to equalize cvs of the stratification variable within the strata. Therefore, this is not the same aim as the aim of stratification, which is to optimize the efficiency of estimation or to minimize a sample size. Furthermore, there is no certainty that under the optimum stratification the distribution of the stratification/survey variable is uniform within the strata. These two sets of strata boundaries (*i.e.*, provided by the geometric and optimization approaches) are not necessarily the same; in fact, they are likely different.

Note that we applied the random search method as the algorithm of the optimization approach to stratification. In fact, Lavallée and Hidirolou's (1988) algorithm is a representative of optimization approaches, too. When the aim of stratification is to minimize a sample size required to achieve a desired level of precision, the two approaches will likely provide similar results, as they did in our study. Nevertheless, the random search algorithm may be applied to any stratification problem (*i.e.*, any optimization function and its constraints), contrary to the LH algorithm, which is applicable only when a sample size is minimized with respect to a given level of precision. It is to be noted that the random search algorithm, as a global optimization method, provides random results.

Our aim, however, was not to promote any of these two algorithms by showing that they are more efficient than the geometric stratification. In addition, we applied Nelder and Mead's (1965) simplex method to stratify the populations (results not presented in the paper); its results were very similar to those of the LH and random search method algorithms. Each of these methods has some drawbacks. For instance, numerical difficulties may occur while using the LH algorithm (Slanta and Krenzke 1996); the random search method provides random results (Kozak 2004); Nelder and Mead's (1965) method may be inefficient under large number of strata and large populations (Kozak 2004); and, in fact, none of the methods has been proven to provide optimum stratification points. Therefore, there is still a need of constructing a stratification algorithm that would be optimum irrespective of the situation (*e.g.*, of a population size or variable's skewness) and that would provide results

that are not random. Our main aim was to prove that the geometric stratification is not optimum, although the stratification points it provides may be useful as initial parameters in other approaches to stratification.

Acknowledgements

The authors are very indebted to the referees and the Associate Editor of *Survey Methodology* for their valuable comments, which helped to improve the first version of this paper.

Appendix

The algorithm given below was proposed by Kozak (2004); we have adapted some of its details to the general stratification problem. In the algorithm, we do not refer to the particular problem of stratification (*i.e.*, we do not define the optimization function and its constraints), since the algorithm works for both problems presented in the paper as well as for other stratification problems. Where required, we refer to "optimization function" (which may be either the variance of an estimator considered or a sample size from a population) and "constraints" (which, depending on the optimization function, may be the constraints (5) and (6), or the constraints (5) combined with the constraint on the level of precision of estimation); certainly, other forms of the optimization function and its constraints may be considered as well.

Let us define a vector \mathbf{a} as follows. It takes values on the interval $(1, N)$, N being the population size. Provided that a population is sorted by the values of a stratification variable X , two elements a_{h-1} and a_h of the vector \mathbf{a} define the stratum h in such a way that this stratum consists of the elements with the index I (which gives the order of an element in the population sorted) that $a_{h-1} < I \leq a_h$, $h = 1, \dots, L$, $a_0 = 0$, $a_L = N$. The algorithm is as follows.

1. Sort the population by the values of the stratification variable.
2. Choose an initial vector \mathbf{a} , *i.e.*, the vector of initial strata boundaries. You may use random integers that satisfy the constraints, but practice shows that better results may be achieved by using approximate strata boundaries obtained via some approximate stratification methods. Calculate the value of the optimization function. Check the constraints; if they are not fulfilled, the initial points have to be changed.

3. For $r = 0, 1, \dots, R$ repeat the following step:
 - a. Generate point \mathbf{a}' by drawing one stratum boundary a_i and changing it as follows

$$\begin{aligned} a'_i &= a_i + j, \\ a'_k &= a_k \quad \text{for } k = 1, \dots, L-1, k \neq i, \end{aligned} \quad (11)$$

where j is the random integer, $j \in \langle -p; -1 \rangle \cup \langle 1; p \rangle$, p being a given integer chosen based on the population size (the larger the population, the larger the p value); usually, it should be between 3 and 5.

- b. Calculate the value of the optimization function.
 - c. If the constraints are satisfied and the value of the optimization function under the vector \mathbf{a}' is smaller than the value under the vector \mathbf{a} , accept the new vector, i.e., $\mathbf{a}_{r+1} = \mathbf{a}'$ (where \mathbf{a}_{r+1} is the vector of strata boundaries in a next iteration); otherwise do not accept the vector, i.e., $\mathbf{a}_{r+1} = \mathbf{a}$.
4. Finish the algorithm if the stopping rule is fulfilled, e.g., if $r = R$, where R is given number of steps, or if in the last m (for instance, 50) steps the value of the optimization function did not change. Finally, calculate the vector \mathbf{k} (the vector of final strata boundaries) on the basis of the values of the vector \mathbf{a} .

References

- Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.
- Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.
- Ekman, G. (1959). An approximation useful in univariate stratification. *Annals of Mathematical Statistics*, 30, 219-229.
- Glasser, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*, 30, 28-32.
- Gunning, P., and Horgan, J.M. (2004). A simple algorithm for stratifying skewed populations. *Survey Methodology*, 30, 159-166.
- Gunning, P., Horgan, J.M. and Yancey, W. (2004). Geometric stratification of accounting data. *J. de Contaduría y Administración*, 214, septiembre-diciembre.
- Hidirolou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.
- Horgan, J.M. (2006). Stratification of skewed populations: A review. *International Statistical Review*, 74(1): 67-76.
- Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5), 797-806.
- Lavallée, P., and Hidirolou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.
- Lednicki, B., and Wieczorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, 6, 287-306.
- Nelder, J.A., and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7, 308-313.
- R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; URL <http://www.R-project.org>.
- Rivest, L.-P. (2002). A generalization of Lavallée and Hidirolou algorithm for stratification in business surveys. *Survey Methodology*, 28, 191-198 (<http://www.mat.ulaval.ca/pages/lpr/>).
- Slanta, J., and Krenzke, T. (1996). Applying the Lavallée and Hidirolou method to obtain stratification boundaries for the Census Bureau's annual Capital Expenditure Survey. *Survey Methodology*, 22, 65-75.

Indirect Sampling: The Foundations of the Generalized Weight Share Method

Jean-Claude Deville and Pierre Lavallée¹

Abstract

To select a survey sample, it happens that one does not have a frame containing the desired collection units, but rather another frame of units linked in a certain way to the list of collection units. It can then be considered to select a sample from the available frame in order to produce an estimate for the desired target population by using the links existing between the two. This can be designated by *Indirect Sampling*.

Estimation for the target population surveyed by Indirect Sampling can constitute a big challenge, in particular if the links between the units of the two are not one-to-one. The problem comes especially from the difficulty to associate a selection probability, or an estimation weight, to the surveyed units of the target population. In order to solve this type of estimation problem, the Generalized Weight Share Method (GWSM) has been developed by Lavallée (1995) and Lavallée (2002). The GWSM provides an estimation weight for every surveyed unit of the target population.

This paper first describes Indirect Sampling, which constitutes the foundations of the GWSM. Second, an overview of the GWSM is given where we formulate the GWSM in a theoretical framework using matrix notation. Third, we present some properties of the GWSM such as unbiasedness and transitivity. Fourth, we consider the special case where the links between the two populations are expressed by indicator variables. Fifth, some special typical linkages are studied to assess their impact on the GWSM. Finally, we consider the problem of optimality. We obtain optimal weights in a weak sense (for specific values of the variable of interest), and conditions for which these weights are also optimal in a strong sense and independent of the variable of interest.

Key Words: Indirect Sampling; Generalized Weight Share Method; Unbiasedness; Optimal Weights.

1. Introduction

To select the samples needed for social or economic surveys, it is useful to have sampling frames, *i.e.*, lists of units intended to provide a way to reach desired target populations. Unfortunately, it happens that one does not have a list containing the desired collection units, but rather another list of units linked in a certain way to the list of collection units. One can speak therefore of two populations U^A and U^B linked to each other, where one wants to produce an estimate for U^B . Unfortunately, a sampling frame is only available for U^A . It can then be considered to select a sample s^A from U^A in order to produce an estimate for U^B by using the correspondence existing between the two populations. This can be designated by *Indirect Sampling*.

Estimation for a target population U^B surveyed by Indirect Sampling can constitute a big challenge, in particular if the links between the units of the two populations are not one-to-one. The problem comes especially from the difficulty to associate a selection probability, or an estimation weight, to the surveyed units of the target population. In order to solve this type of estimation problem, the Generalized Weight Share Method (GWSM) has been developed by Lavallée (1995) and Lavallée (2002), and presented also in Lavallée and Caron (2001). The

GWSM provides an estimation weight for every surveyed unit of the target population U^B . Basically, this estimation weight corresponds to a weighted average of the survey weights of the units of the sample s^A . The GWSM is an extension of the Weight Share Method described by Ernst (1989) in the context of longitudinal household surveys.

The purposes of this paper are to describe Indirect Sampling—the foundations underlying the GWSM—and to obtain optimal weights from the GWSM that provide unbiased estimates with minimum variance. First, we will describe Indirect Sampling together with the GWSM in a theoretical framework that will use, for instance, matrix notation. The use of matrix notation for the GWSM has previously been presented by Deville (1998). Second, we will use this theoretical framework to state some general properties associated with the GWSM that include unbiasedness and transitivity. Transitivity is to go from the population U^A to a target population U^C , through an intermediate population U^B . Third, we will show the correspondence between the matrix formulation and the one that has been described in Lavallée (1995), Lavallée (2002), and Lavallée and Caron (2001). Fourth, we will study the effect of various typical link matrices between U^A and U^B on the precision of the estimates obtained from the GWSM. Finally, we will assess the problem of optimality. We will obtain optimal weights in a weak sense (for specific values

1. Jean-Claude Deville, Laboratoire de Statistique d'Enquête (ENSAI/CREST), Campus de Ker Lann, rue Blaise Pascal, 35170 Bruz, FRANCE. E-mail: deville@ensai.fr; Pierre Lavallée, Statistics Canada, Ottawa, Ontario, K1A 0T6, CANADA. E-mail: pierre.lavallee@statcan.ca.

of the variable of interest), and conditions under which these weights are also optimal in a strong sense and independent of the variable of interest.

2. Indirect Sampling

As mentioned in the introduction, with Indirect Sampling, we select a sample s^A from a population U^A in order to produce an estimate for a target population U^B . For that, we use the correspondence existing between the two populations. For example, assume that we want to produce estimates for a population of children (collection units) while we only have a sampling frame of parents. The target population U^B is the one of the children, but we need to select a sample of parents before being able to interview the children. This is illustrated in Figure 1.

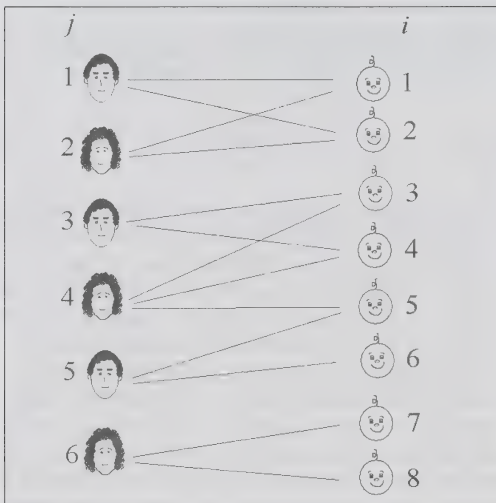


Figure 1. Population U^A of parents and population U^B of children with the links between the two.

Let the population U^A contain N^A units, where each unit is labeled by the letter j . Similarly, let the target population U^B contain N^B units, where each unit is labeled by the letter i . The correspondence between the two populations U^A and U^B can be represented by a *link matrix* $\Theta_{AB} = [\theta_{ji}^{AB}]$ of size $N^A \times N^B$ where each element $\theta_{ji}^{AB} \geq 0$. That is, unit j of U^A is related to unit i of U^B provided that $\theta_{ji}^{AB} > 0$, otherwise the two units are not related to each other. For the above example, the link matrix is given by

$$\Theta_{AB} = \begin{bmatrix} \theta_{11}^{AB} & \theta_{12}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{21}^{AB} & \theta_{22}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{33}^{AB} & \theta_{34}^{AB} & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{43}^{AB} & \theta_{44}^{AB} & \theta_{45}^{AB} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{55}^{AB} & \theta_{56}^{AB} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta_{67}^{AB} & \theta_{68}^{AB} \end{bmatrix}.$$

Obtaining the link matrix *link matrix* $\Theta_{AB} = [\theta_{ji}^{AB}]$ is a critical issue in Indirect Sampling. For the case where two units $j \in U^A$ and $i \in U^B$ are not linked, we simply set $\theta_{ji}^{AB} = 0$. When there is a link between two units j and i , the choice of $\theta_{ji}^{AB} > 0$ is important. As we will see, it influences the precision of the estimates issued from Indirect Sampling. Now, in several applications, the values of θ_{ji}^{AB} for the linked units are simply set to 1. Of course, the values of θ_{ji}^{AB} for the linked units can be chosen to be different from 1. Lavallée and Caron (2001) discussed the use of the linkage weights obtained from a record linkage process between U^A and U^B for assigning values to the θ_{ji}^{AB} . The linkage weights are proportional to the probability of two units $j \in U^A$ and $i \in U^B$ being linked. Since the choice of $\theta_{ji}^{AB} > 0$ for two linked units j and i can affect the precision of the estimates, it is natural to seek for those θ_{ji}^{AB} that will minimize the variance of the estimates. This optimization problem is considered in section 6 of the paper.

With Indirect Sampling, we select the sample s^A of n^A units from U^A using some sampling design. Let π_j^A be the selection probability of unit j . We assume $\pi_j^A > 0$ for all $j \in U^A$. For each unit j selected in s^A , we identify the units i of U^B that have a non-zero correspondence, i.e., with $\theta_{ji}^{AB} > 0$. Let Ω^B be the set of the n^B units of U^B identified by the units $j \in s^A$, i.e., $\Omega^B = \{i \in U^B \mid \exists j \in s^A \text{ and } \theta_{ji}^{AB} > 0\}$. For each unit i of the set Ω^B , we measure a variable of interest y_i from the target population U^B . Let $\mathbf{Y} = \{y_1, \dots, y_{N^B}\}'$ be the column vector of that variable of interest. In a practical view point, it is important to mention that although the sample size n^A is usually determined in advance, the number of units n^B is difficult to control because it depends on the selected sample s^A and the link matrix Θ_{AB} . As a consequence, it turns out to be difficult in general to establish a budget for measuring the variable of interest y_i . Fortunately, in most applications (e.g., the parents-children case above), the number of links that start from a given unit j of s^A is somewhat predictable (for example, a parent typically has one, two, or three children), which helps to assess how many units i of U^B will finally be measured.

We assume that for any unit j of s^A , the correspondences for $i = 1, \dots, N^B$ can be obtained. That is, we can identify all the links between the two populations by direct interview or by some administrative source for any sampled

unit j . Also, for any identified unit i of U^B , we assume that the links for $j = 1, \dots, N^A$ can be obtained (as mentioned by Lavallée (2002), there are cases where this last constraint can be difficult to satisfy in practice. Referring to the example of parents and children, it might not be easy for a very young child, selected through his mother, to mention back his father, when the two parents are divorced. In order to simplify the discussion, such a problem of identification of links will be assumed to be negligible). Therefore, the values of the links need not to be known between the entire populations U^A and U^B . In fact, we need to know the links (and consequently the values of θ_{ji}^{AB}) only for the lines j of Θ_{AB} where $j \in s^A$, and also for columns i of Θ_{AB} where $i \in \Omega^B$.

Suppose that we are interested in estimating the total Y^B of the target population U^B where $Y^B = \sum_{i=1}^{N^B} y_i$. We can also write $Y^B = \mathbf{1}'_B \mathbf{Y}$ where $\mathbf{1}_B$ is the column vector of 1's of size N^B (note that we use for simplification the notation $\mathbf{1}_B$ instead of $\mathbf{1}_{N^B}$). Now let $\theta_{+i}^{AB} = \sum_{j=1}^{N^A} \theta_{ji}^{AB}$ and let $\theta_{ji}^{AB} = \theta_{ji}^{AB} / \theta_{+i}^{AB}$. We have $\mathbf{1}'_A \Theta_{AB} = \{\theta_{+1}^{AB}, \dots, \theta_{+N^B}^{AB}\}$. We then define the *standardized link matrix* $\tilde{\Theta}_{AB} = \Theta_{AB} [\text{diag}(\mathbf{1}'_A \Theta_{AB})]^{-1}$, where $\text{diag}(\mathbf{v})$ is the square matrix obtained by putting the elements of the row-vector (or column-vector) \mathbf{v} in the diagonal, and 0 elsewhere. Note that in order for the matrix $\tilde{\Theta}_{AB}$ to be well defined, we must have $[\text{diag}(\mathbf{1}'_A \Theta_{AB})]^{-1}$ to exist, which is the case if and only if $\theta_{+i}^{AB} > 0$ for all $i = 1, \dots, N^B$. For the parents-children example, this means that every child must be linked to at least a parent.

Result 1:

The link matrix $\tilde{\Theta}_{AB}$ is a standardized link matrix if and only if

$$\tilde{\Theta}_{AB} \mathbf{1}_A = \mathbf{1}_B. \quad (2.1)$$

The proof of Result 1 follows directly from the definition of a standardized link matrix. Using Result 1, we directly obtain Result 2 that can also be found in Deville (1998):

Result 2:

$$\begin{aligned} Y^B &= \mathbf{1}'_B \mathbf{Y} \\ &= \mathbf{1}'_A \tilde{\Theta}_{AB} \mathbf{Y} = \sum_{j=1}^{N^A} \sum_{i=1}^{N^B} \theta_{ji}^{AB} y_i. \end{aligned} \quad (2.2)$$

Let us define the column vector $\mathbf{Z} = \tilde{\Theta}_{AB} \mathbf{Y}$ of size N^A . Considering each line of \mathbf{Z} , the variable $z_j = \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB} y_i$ is defined for each unit j of the population U^A and measured for each unit $j \in s^A$.

For estimating Y^B , we want to use the values of y_i measured from set Ω^B . For this, we will use an estimator of the form:

$$\hat{Y}^B = \sum_{i=1}^{N^B} w_i y_i \quad (2.3)$$

where w_i is the estimation weight of the unit i of Ω^B , with $w_i = 0$ for $i \notin \Omega^B$. Let $\mathbf{W}' = \{w_1, \dots, w_{N^B}\}$. The estimator (2.3) can be rewritten as

$$\hat{Y}^B = \mathbf{W}' \mathbf{Y}. \quad (2.4)$$

Usually, to get an unbiased estimate of Y^B , one can simply use as the weight the inverse of the selection probability π_i^B of unit i . As mentioned by Lavallée (1995) and Lavallée (2002), with Indirect Sampling, this probability can however be difficult, or even impossible, to obtain. It is then proposed to use the GWSM, which is defined as follows.

Let $\boldsymbol{\pi}^A = \{\pi_1^A, \dots, \pi_{N^A}^A\}'$ and let $\Pi_A = \text{diag}(\boldsymbol{\pi}^A)$ be the diagonal matrix of size $N^A \times N^A$ containing the selection probabilities used for the selection of sample s^A . Accordingly, let $\mathbf{t}^A = \{t_1^A, \dots, t_{N^A}^A\}'$ where $t_j^A = 1$ if $j \in s^A$, and 0 otherwise. Let $\mathbf{T}_A = \text{diag}(\mathbf{t}^A)$ be the diagonal matrix of size $N^A \times N^A$ containing the indicator variables t_j^A . Starting from $Y^B = \mathbf{1}'_A \tilde{\Theta}_{AB} \mathbf{Y} = \mathbf{1}'_A \mathbf{Z}$, we can directly form the following Horvitz-Thompson estimator in terms of the vector \mathbf{Z} :

$$\hat{Y}^B = \mathbf{1}'_A \mathbf{T}_A \Pi_A^{-1} \mathbf{Z}. \quad (2.5)$$

Using the fact that $\mathbf{Z} = \tilde{\Theta}_{AB} \mathbf{Y}$, we have $\hat{Y}^B = \mathbf{1}'_A \mathbf{T}_A \Pi_A^{-1} \tilde{\Theta}_{AB} \mathbf{Y}$ and therefore we can define the column vector \mathbf{W} of weights:

$$\mathbf{W} = \tilde{\Theta}_{AB} \mathbf{T}_A \Pi_A^{-1} \mathbf{1}_A. \quad (2.6)$$

The vector \mathbf{W} is of size N^B and for each $i = 1, \dots, N^B$, we have $w_i = \sum_{j=1}^{N^A} t_j^A \tilde{\theta}_{ji}^{AB} / \pi_j^A$. The weights w_i of that vector are said to be obtained from the GWSM, as described by Lavallée (2002).

3. Properties of the GWSM

3.1 Unbiasedness

As mentioned by Ernst (1989), to get an unbiased estimator, we only need to have $E(\mathbf{W}) = \mathbf{1}_B$. By construction, because the estimator (2.5) is a Horvitz-Thompson estimator, this condition is directly satisfied and therefore, the GWSM produces unbiased estimates.

From this discussion, we can in addition obtain the following result:

Result 3:

The vector of weights \mathbf{W} given by (2.6) provides unbiased estimates if and only if the matrix $\tilde{\Theta}_{AB}$ is a standardized link matrix.

Proof:

Starting from (2.6), we have

$$E(\mathbf{W}) = \tilde{\Theta}'_{AB} \mathbf{1}_A. \quad (3.1)$$

Using Result 1, we directly get $E(\mathbf{W}) = \mathbf{1}_B$ and therefore we have unbiased estimates. Now, assume that $E(\mathbf{W}) = \mathbf{1}_B$. From (3.1), we must have $\tilde{\Theta}'_{AB} \mathbf{1}_A = \mathbf{1}_B$ and therefore, $\tilde{\Theta}_{AB}$ is a standardized link matrix.

3.2 Variance

Because the estimator (2.5) is a Horvitz-Thompson estimator, we directly obtain the following result:

Result 4:

The variance of \hat{Y}^B is given by

$$\begin{aligned} \text{Var}(\hat{Y}^B) &= \mathbf{Z}' \Delta_A \mathbf{Z} \\ &= \mathbf{Y}' \Delta_B \mathbf{Y} \end{aligned} \quad (3.2)$$

where $\Delta_A = [(\pi_{ji}^A - \pi_j^A \pi_{j'}^A) / \pi_j^A \pi_{j'}^A]_{N^A \times N^A}$ is a non-negative definite matrix of size $N^A \times N^A$ and where $\pi_{jj'}^A$ is the joint selection probability of units j and j' from U^A , and where $\Delta_B = \tilde{\Theta}'_{AB} \Delta_A \tilde{\Theta}_{AB}$.

For a proof of the variance of the Horvitz-Thompson estimator, see Särndal, Swensson and Wretman (1992).

3.3 Transitivity

Let us suppose that we are interested in producing estimates for a target population U^C that can only be reached through the population U^B . We assume that the target population U^C contains N^C units, where each unit is labeled by the letter k . The correspondence between the two populations U^B and U^C can be represented by the link matrix $\Theta_{BC} = [\theta_{ik}^{BC}]$ of size $N^B \times N^C$ where each element $\theta_{ik}^{BC} \geq 0$. That is, unit i of U^B is related to unit k of U^C provided that $\theta_{ik}^{BC} > 0$, otherwise the two units are not related to each other.

We can now use Indirect Sampling by *transitivity*. For this, we select a sample s^A from the population U^A and first identify the set Ω^B of U^B . From this set Ω^B , we then identify the units of U^C that are associated in order to form the set $\Omega^C = \{k \in U^C \mid \exists i \in \Omega^B \text{ and } \theta_{ik}^{BC} > 0\}$ of units to be measured from the target population U^C . An important question is to see if the GWSM, when applied in the context of Indirect Sampling by transitivity, is also transitive. That is, is applying the GWSM from U^A to U^B , and then from U^B to U^C , is equivalent to directly applying the GWSM from U^A to U^C ?

First, consider using Indirect Sampling from U^A directly to the target population U^C . By going from the population U^A to U^B , and then to U^C , this can relate to having the link matrix $\Theta_{AC} = [\theta_{jk}^{AC}]$ of size $N^A \times N^C$ defined as $\Theta_{AC} = \Theta_{AB} \Theta_{BC}$. For each unit j selected in s^A , we identify the

units k of U^C that have a non-zero correspondence, i.e., with $\theta_{jk}^{AC} > 0$, to obtain the set $\bar{\Omega}^C = \{k \in U^C \mid \exists j \in s^A \text{ and } \theta_{jk}^{AC} > 0\}$. We measure the variable of interest y_k from the target population U^C . Applying the GWSM, we obtain from (2.6) the following weights:

$$\bar{\mathbf{W}}_C = \tilde{\Theta}'_{AC} \mathbf{T}_A \Pi_A^{-1} \mathbf{1}_A \quad (3.3)$$

where $\tilde{\Theta}_{AC} = \Theta_{AC} [\text{diag}(\mathbf{1}'_A \Theta_{AC})]^{-1}$.

Let us now consider using Indirect Sampling in two steps. For each unit j selected in s^A , we identify the units i of U^B that have a non-zero correspondence, i.e., with $\theta_{ji}^{AB} > 0$. As before, we have $\Omega^B = \{i \in U^B \mid \exists j \in s^A \text{ and } \theta_{ji}^{AB} > 0\}$. For each unit i of the set Ω^B , we then identify the units k of U^C that have a non-zero correspondence, i.e., with $\theta_{ik}^{BC} > 0$. We then have the set $\Omega^C = \{k \in U^C \mid \exists i \in \Omega^B \text{ and } \theta_{ik}^{BC} > 0\}$. From (2.6), we have the column vector \mathbf{W}_B of weights associated to the units of population U^B :

$$\mathbf{W}_B = \tilde{\Theta}'_{AB} \mathbf{T}_A \Pi_A^{-1} \mathbf{1}_A. \quad (3.4)$$

For each unit i of the set Ω^B , we then have a non-zero weight w_i^B . Now, the set Ω^B can be seen as a sample of units that are used in an Indirect Sampling process to identify the set Ω^C . By similarity with Indirect Sampling from the sample s^A to the target population U^B , applying the GWSM in the context of Indirect Sampling from the set Ω^B to the target population U^C produces the following weights:

$$\mathbf{W}_C = \tilde{\Theta}'_{BC} \mathbf{T}_B \text{diag}(\mathbf{W}_B) \mathbf{1}_B \quad (3.5)$$

where $\tilde{\Theta}_{BC} = \Theta_{BC} [\text{diag}(\mathbf{1}'_B \Theta_{BC})]^{-1}$ and $\mathbf{T}_B = \text{diag}(\mathbf{t}_B)$ with $\mathbf{t}_B = (t_1^B, \dots, t_{N^B}^B)'$ and $t_i^B = 1$ if $i \in \Omega^B$, and 0 otherwise. Because the weights $w_i^B = 0$ for $i \notin \Omega^B$, we have $\mathbf{T}_B \text{diag}(\mathbf{W}_B) = \text{diag}(\mathbf{W}_B)$. Therefore, we obtain

$$\mathbf{W}_C = \tilde{\Theta}'_{BC} \text{diag}(\mathbf{W}_B) \mathbf{1}_B. \quad (3.6)$$

Replacing \mathbf{W}_B by (3.4) in equation (3.6), we get

$$\begin{aligned} \mathbf{W}_C &= \tilde{\Theta}'_{BC} \text{diag}(\tilde{\Theta}'_{AB} \mathbf{T}_A \Pi_A^{-1} \mathbf{1}_A) \mathbf{1}_B \\ &= \tilde{\Theta}'_{BC} \tilde{\Theta}'_{AB} \mathbf{T}_A \Pi_A^{-1} \mathbf{1}_A. \end{aligned} \quad (3.7)$$

Since $\tilde{\Theta}'_{BC} \tilde{\Theta}'_{AB} \mathbf{1}_A = \tilde{\Theta}'_{BC} \mathbf{1}_B = \mathbf{1}_C$, from Result 1, the matrix $\tilde{\Theta}_{AB} \tilde{\Theta}_{BC}$ is a standardized link matrix. Because of this, the GWSM is therefore transitive, at least in some sense. That is, the weights \mathbf{W}_C can be obtained in a single step by using the standardized link matrix $\tilde{\Theta}_{AB} \tilde{\Theta}_{BC}$ into the GWSM. Now, for the GWSM to be perfectly transitive, the weights \mathbf{W}_C provided (3.7) would need to be exactly the same as the weights $\bar{\mathbf{W}}_C$ provided by (3.3). By comparing equations (3.3) and (3.7), we obtain the following result:

Result 5:

Applying the GWSM from U^A to U^B , and then from U^B to U^C , is transitive if and only if

$$\tilde{\Theta}_{AC} = \tilde{\Theta}_{AB} \tilde{\Theta}_{BC} \quad (3.8)$$

Unfortunately, condition (3.8) does not hold in general. In fact, it is relatively easy to construct examples where $\tilde{\Theta}_{AC} \neq \tilde{\Theta}_{AB} \tilde{\Theta}_{BC}$.

4. A Structural Property of the GWSM

In the present section, we stress the fact that with Indirect Sampling, the sampling process depends only on the links between the two populations U^A and U^B . The values of the θ_{ji}^{AB} themselves, apart from being zero or not, do not interfere in the sampling process. On the other hand, the values of the θ_{ji}^{AB} **do** have a role in the weights, and therefore the estimator, issued from the GWSM. We extend this idea in the following paragraphs.

Indirect Sampling associates to each sample s^A in U^A a sample Ω^B in U^B , namely $\Omega^B = \{i \in U^B \mid \exists j \in s^A \text{ and } \theta_{ji}^{AB} > 0\}$. Thus, a function $f: s^A \rightarrow \Omega^B$ that maps the sample s^A to the sample Ω^B is uniquely determined by the set of couples (j, i) with $\theta_{ji}^{AB} > 0$. Let $I_{ji}^{AB} = 1$ if $\theta_{ji}^{AB} > 0$, and 0 otherwise. These are the elements of the incidence matrix of the graph linking U^A to U^B .

Suppose we are given a function ϕ from the set of subsets of U^A into the set of subsets of U^B . Like f , suppose that ϕ satisfies the "Union Property": $\phi(s_1^A \cup s_2^A) = \phi(s_1^A) \cup \phi(s_2^A)$, where s_1^A and s_2^A are two subsets of U^A .

Result 6:

The function ϕ is determined unequivocally by a *zero-one link matrix*.

Proof:

This can be shown as follows: Take $s_j^A = \{j\}$ for some unit j in U^A . Then, $\phi(s_j^A)$ is a set in U^B . Let $I_{ji}^{AB} = 1$ if unit i of U^B belongs to $\phi(s_j^A)$, and 0 otherwise. By the Union Property, $\phi(s^A) = \bigcup_{j \in s^A} \phi(s_j^A)$ and the set of I_{ji}^{AB} defines the *zero-one link matrix* $\mathbf{L}_{AB} = [I_{ji}^{AB}]$ of size $N^A \times N^B$, which precisely defines the function ϕ .

This provides us an equivalence relation between link matrices, associated with a deeper property. Let p^A be a sampling design on U^A (i.e., a probability distribution on the set of subsets of U^A). The function f induces a sampling design on U^B by $p^B(\Omega^B) = \sum_{s^A: \Omega^B = f(s^A)} p^A(s^A)$. As the design is induced by f , it does not depend on the particular link matrix Θ_{AB} defining the function, but is rather a characteristic of the equivalence class through the zero-one link matrix \mathbf{L}_{AB} . As a consequence, the Horvitz-Thompson estimator in U^B depends only on this class. It is therefore of some interest to choose in this class a matrix

Θ_{AB} having, in some sense, an optimal characteristic (see section 6).

5. Special Link matrices

As it can be seen from the previous sections, the link matrix Θ_{AB} drives the form of the estimator (2.4) obtained from the GWSM. In this section, we present some special link matrices Θ_{AB} that correspond to extreme cases. Although not all such cases are likely to be seen in practice, they illustrate the effect of the link matrix on the estimator (2.4).

5.1 Identity Matrix

Assume that the link matrix Θ_{AB} is given by the identity matrix \mathbf{I} . In practice, this means that the population U^A and the target population U^B have a one-to-one relationship. Of course, this implies that $N^A = N^B = N$ and that the identity matrix \mathbf{I} is of size $N \times N$.

As a first result, we have $\tilde{\Theta}_{AB} = \mathbf{I}$. As a consequence, the vector of weights (2.6) is given by $\mathbf{W}' = (t_1^A / \pi_1^A, \dots, t_{N^A}^A / \pi_{N^A}^A)$ and we also have $\mathbf{Z} = \tilde{\Theta}_{AB} \mathbf{Y} = \mathbf{Y}$. Therefore, the estimator \hat{Y}^B given by (2.5) turns out to be nothing else than the Horvitz-Thompson estimator $\hat{Y}^B = \mathbf{1}'_A \mathbf{T}_A \mathbf{\Pi}_A^{-1} \mathbf{Y}$.

5.2 One for All (Within Clusters)

Consider the case where the population U^B is divided into Γ clusters where each cluster γ is of size N_γ^B . These clusters are such that each cluster γ from U^B is associated to exactly one unit j of U^A . Because of this, we can use the letter γ for both the units j from U^A and the clusters from U^B . Note also that $\Gamma = N^A$.

This situation corresponds to a link matrix Θ_{AB} being block diagonal where each submatrix contains only one line. Let the row vector $\mathbf{1}'_{B\gamma}$ be of size N_γ^B and containing only 1's. The link matrix Θ_{AB} is then defined as

$$\Theta_{AB} = \begin{bmatrix} \mathbf{1}'_{B1} & & 0 & \dots & 0 \\ & \ddots & & & \\ 0 & & \mathbf{1}'_{B\gamma} & & 0 \\ \vdots & \ddots & & \ddots & \\ 0 & \dots & 0 & & \mathbf{1}'_{B\Gamma} \end{bmatrix} \quad (5.1)$$

We can also write $\Theta_{AB} = \text{diag}(\{\mathbf{1}'_{B1}, \dots, \mathbf{1}'_{B\Gamma}\})$. Using this, we have $\text{diag}(\mathbf{1}'_A \Theta_{AB}) = \text{diag}(\mathbf{1}'_A \text{diag}(\{\mathbf{1}'_{B1}, \dots, \mathbf{1}'_{B\Gamma}\})) = \text{diag}(\{\mathbf{1}'_{B1}, \dots, \mathbf{1}'_{B\Gamma}\})$ and hence $\tilde{\Theta}_{AB} = \Theta_{AB}$. From equation (2.6), we obtain the column vector of weights $\mathbf{W}' = (t_1^A / \pi_1^A \mathbf{1}'_{B1}, \dots, t_\Gamma^A / \pi_\Gamma^A \mathbf{1}'_{B\Gamma})$. As we can see, the elements of the column vector \mathbf{W} have the values $t_\gamma^A / \pi_\gamma^A$ repeated within each cluster γ of U^B . From (2.4), we obtain

$$\hat{Y}^B = \sum_{\gamma=1}^{\Gamma} \frac{t_{\gamma}^A}{\pi_{\gamma}^A} Y_{\gamma}^B \quad (5.2)$$

where $Y_{\gamma}^B = \sum_{i=1}^{N_{\gamma}^B} y_{i\gamma}$.

5.3 All for One (Within Clusters)

Consider the case where the population U^A is divided into Γ clusters where each cluster γ is of size N_{γ}^A . These clusters are such that each cluster γ from U^A is associated to exactly one unit i of U^B . Because of this, we can use the letter γ for both the clusters from U^A and the units i from U^B . Note also that $\Gamma = N^B$.

This situation corresponds to a link matrix Θ_{AB} being block diagonal where each submatrix contains only one column. Let the column vector $\mathbf{1}_{A\gamma}$ be of size N_{γ}^A and containing only 1's. The link matrix Θ_{AB} is then defined as

$$\Theta_{AB} = \begin{bmatrix} \mathbf{1}_{A1} & \mathbf{0} & \cdots & \mathbf{0} \\ & \ddots & & \vdots \\ \mathbf{0} & \mathbf{1}_{A\gamma} & & \mathbf{0} \\ & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1}_{A\Gamma} \end{bmatrix}. \quad (5.3)$$

We can also write $\Theta_{AB} = \text{diag}(\{\mathbf{1}_{A1}, \dots, \mathbf{1}_{A\Gamma}\})$. Using this, we have $\tilde{\Theta}_{AB} = \text{diag}(\{1/N_1^A \mathbf{1}_{A1}, \dots, 1/N_{\Gamma}^A \mathbf{1}_{A\Gamma}\})$. From equation (2.6), we obtain the column vector of weights $\mathbf{W}' = (1/N_1^A \sum_{j=1}^{N_1^A} t_j^A / \pi_j^A, \dots, 1/N_{\Gamma}^A \sum_{j=1}^{N_{\Gamma}^A} t_j^A / \pi_j^A)$. Thus, the elements γ (or i) of the column vector \mathbf{W} have the averaged values $\sum_{j=1}^{N_{\gamma}^A} t_j^A / \pi_j^A N_{\gamma}^A$, $\gamma = 1, \dots, \Gamma$. From (2.4), we obtain $\hat{Y}^B = \sum_{\gamma=1}^{\Gamma} y_{\gamma} / N_{\gamma}^A \sum_{j=1}^{N_{\gamma}^A} t_j^A / \pi_j^A$.

5.4 Inefficient Sampling

Suppose that some rows of the link matrix Θ_{AB} contain only zeros. This means that some units of the population U^A are not associated to any unit of the target population U^B . Then, if such units are selected in the sample s^A , this will lead to the identification of no unit from U^B . This can be seen as inefficient in a sampling point of view. In a more formal way, assume that each of the first N^{1A} rows of the link matrix Θ_{AB} contains at least one $\theta_{ji} > 0$, and that they form the submatrix Θ_1 . Assume that the other N^{0A} rows of Θ_{AB} have $\theta_{ji} = 0$ for $i = 1, \dots, N^B$. We therefore have

$$\Theta_{AB} = \begin{bmatrix} \Theta_1 \\ \mathbf{0} \end{bmatrix}.$$

As a first result, we obtain

$$\tilde{\Theta}_{AB} = \begin{bmatrix} \Theta_1 [\text{diag}(\mathbf{1}'_A \Theta_1)]^{-1} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \tilde{\Theta}_1 \\ \mathbf{0} \end{bmatrix} \quad (5.4)$$

where $\mathbf{1}_{1A}$ is the column vector of 1's of size N^{1A} . From equation (2.6), we obtain the column vector of weights $\mathbf{W} = [\tilde{\Theta}_1' \mathbf{0}'] \mathbf{T}_A \Pi_{1A}^{-1} \mathbf{1}_{1A}$. Let $\Pi_{1A} = \text{diag}(\{\pi_1^A, \dots, \pi_{N^{1A}}^A\})$ be the diagonal matrix of size $N^{1A} \times N^{1A}$ and accordingly, let $\mathbf{T}_{1A} = \text{diag}(\{t_1^A, \dots, t_{N^{1A}}^A\})$ be the diagonal matrix of size $N^{1A} \times N^{1A}$. We then get

$$\begin{aligned} \mathbf{W} &= [\tilde{\Theta}_1' \mathbf{0}'] \mathbf{T}_A \Pi_{1A}^{-1} \mathbf{1}_{1A} \\ &= \tilde{\Theta}_1' \mathbf{T}_{1A} \Pi_{1A}^{-1} \mathbf{1}_{1A}. \end{aligned} \quad (5.5)$$

As we can see from (5.5), the weights only depend on the probabilities of selection π_j^A of the units of U^A that have at least one $\theta_{ji} > 0$ for $i = 1, \dots, N^B$. From (2.4), we finally obtain $\hat{Y}^B = \mathbf{1}'_{1A} \mathbf{T}_{1A} \Pi_{1A}^{-1} \tilde{\Theta}_1 \mathbf{Y}$.

5.5 Biased Estimator

Suppose that some columns of the link matrix Θ_{AB} contain only zeros. This means that some units of the population U^B are not associated to any unit of the target population U^A . Recall that in order for the matrix $\tilde{\Theta}_{AB}$ to be well defined, we must have $\text{diag}(\mathbf{1}'_A \Theta_{AB})^{-1}$ to exist. As we will see, the present case does not satisfy this condition. This results in a biased estimator for the total Y^B .

In a more formal way, assume that each of the first N^{1B} columns of the link matrix Θ_{AB} contains at least one $\theta_{ji} > 0$, and let them form the submatrix Θ_1 , different from the one of the previous section. Assume that the other N^{0B} columns of Θ_{AB} have $\theta_{ji} = 0$ for $j = 1, \dots, N^A$. We therefore have $\Theta_{AB} = [\Theta_1, \mathbf{0}]$.

From this definition, we directly have

$$\begin{aligned} [\text{diag}(\mathbf{1}'_A \Theta_{AB})]^{-1} &= [\text{diag}(\mathbf{1}'_A \Theta_1, \mathbf{1}'_A \mathbf{0})]^{-1} \\ &= \begin{bmatrix} \text{diag}(\mathbf{1}'_A \Theta_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^{-1}. \end{aligned} \quad (5.6)$$

Since this matrix is singular, $[\text{diag}(\mathbf{1}'_A \Theta_{AB})]^{-1}$ does not exist. As a solution to this problem, it could be possible to use a *generalized inverse*. Recall that for a given square matrix \mathbf{A} , the matrix \mathbf{A}^- is a generalized inverse of \mathbf{A} provided that $\mathbf{A} \mathbf{A}^- \mathbf{A} = \mathbf{A}$ (Searle 1971). One possible generalized inverse of (5.6) is

$$[\text{diag}(\mathbf{1}'_A \Theta_{AB})]^- = \begin{bmatrix} [\text{diag}(\mathbf{1}'_A \Theta_1)]^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \quad (5.7)$$

With this generalized inverse, we have the following standardized link matrix $\tilde{\Theta}_- = \Theta_{AB} [\text{diag}(\mathbf{1}'_A \Theta_{AB})]^- = [\tilde{\Theta}_1, \mathbf{0}]$. Starting from equation (2.6), we can obtain the column vector \mathbf{W}_- of weights:

$$\mathbf{W}_- = \begin{bmatrix} \tilde{\Theta}_1' \mathbf{T}_A \Pi_{1A}^{-1} \mathbf{1}_{1A} \\ \mathbf{0}' \end{bmatrix}. \quad (5.8)$$

As we can see from (5.8), the weights are null for the units i of the target population U^B that Θ_{AB} have $\theta_{ji} = 0$ for $j = 1, \dots, N^A$. From (2.4) and using \mathbf{W}_- instead of \mathbf{W} , we obtain $\hat{Y}^B = \mathbf{1}_A \mathbf{T}_A \Pi_A^{-1} \tilde{\Theta} \mathbf{Y}_1$ where $\mathbf{Y}_1 = \{y_1, \dots, y_{N^{1B}}\}'$ is the subvector constructed from the N^{1B} first elements of \mathbf{Y} . Since in general $E(\hat{Y}^B) = \mathbf{1}_A' \tilde{\Theta} \mathbf{1}_1 \mathbf{Y}_1 \neq \mathbf{1}_A' \mathbf{Y} = Y^B$, this estimator is biased for the total Y^B .

6. Optimality

Optimality is an important aspect of the GWSM. As it has been shown in Result 3, the estimator \hat{Y}^B obtained by the GWSM will provide unbiased estimates provided that the matrix $\tilde{\Theta}_{AB}$ is a standardized link matrix. Now, given that the variance (3.2) of this estimator depends on this matrix, there should be at least one matrix $\tilde{\Theta}_{AB, \text{opt}}$ such that the variance of the estimator \hat{Y}^B will be minimum. That is, for the θ_{ji}^{AB} that are greater than 0, we are interested in finding the values that these θ_{ji}^{AB} should have to obtain the most precise estimator \hat{Y}^B .

This optimality problem was first assessed by Kalton and Brick (1995). They obtained results based on the simplified situation where $N^A = 2$ and with s^A obtained through equal probability sampling. Their conclusions suggested the use of $\theta_{ji}^{AB, \text{opt}} = 1$ when $\theta_{ji}^{AB} > 0$, and $\theta_{ji}^{AB, \text{opt}} = 0$ when $\theta_{ji}^{AB} = 0$. Lavallée (2002) and Lavallée and Caron (2001) obtained results along the same lines by the use of simulations. In the present section, we present new results on the optimality of the GWSM.

6.1 Factorization

Factorization is the reverse problem of transitivity. It consists in finding a population U^G and standardized link matrices $\tilde{\Theta}_{AG}$ and $\tilde{\Theta}_{GB}$ such that $\tilde{\Theta}_{AB} = \tilde{\Theta}_{AG} \tilde{\Theta}_{GB}$. This leads to an important simplification in searching for an optimal standardized link matrix $\tilde{\Theta}_{AB, \text{opt}}$.

The population U^G can be taken as being one of clusters, the factorization being achieved in the context of “one for all (within clusters)” (from U^A to U^G) and “all for one (within clusters)” (from U^G to U^B), as presented in sections 5.2 and 5.3. This can be described in a very general way as follows. Consider a population U^G containing as many units as there are links starting from the units j of U^A . The population size N^G is then given by the number of θ_{ji}^{AB} of $\tilde{\Theta}_{AB}$ that are greater than 0. Each unit g of U^G can be seen as the extremity of an “arrow” starting from some unit j of U^A . From this graph, there is only one link matrix $\tilde{\Theta}_{AG}$ of size $N^A \times N^G$ keeping unbiasedness, namely $\tilde{\Theta}_{AG} = [\theta_{jg}^{AG}]$ where $\theta_{jg}^{AG} = 1$ if there is a link (or an “arrow”) leaving unit j of U^A to unit g from U^G , and $\theta_{jg}^{AG} = 0$ otherwise. Note that by construction, each unit g

from U^G is linked to at most one unit j from U^A and therefore $\tilde{\Theta}_{AG} = \Theta_{AG}$. This corresponds to the “one to all within clusters” situation presented in section 5.2. Indirect Sampling from U^A to U^G is in fact standard Cluster Sampling and leading the GWSM to the usual Horvitz-Thompson estimator (see Lavallée 2002). For the parent-child example, the result of this factorization would be given by Figure 2.

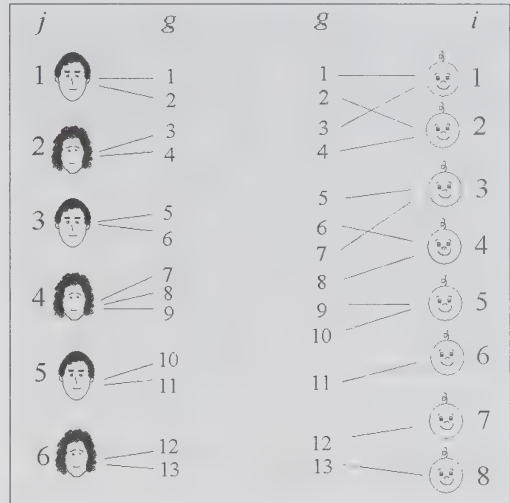


Figure 2. Result of the factorization of the parents-children populations.

Considering the graph from U^G to U^B , we can construct the link matrix $\tilde{\Theta}_{GB}$ of size $N^G \times N^B$ as follows. Because of the definition of the population U^G , each unit g of U^G is linked to exactly one unit i of U^B . Note that Indirect Sampling in this context can be seen as sampling clusters (i.e., the units i of U^B) from their elements (i.e., the units g of U^G). It can also be seen as the “all to one within clusters” presented in section 5.3. Let $\tilde{\Theta}_{GB} = \Theta_{GB} [\text{diag}(\mathbf{1}_G' \Theta_{GB})]^{-1}$ be the standardized link matrix obtained from Θ_{GB} . We have $\text{diag}(\mathbf{1}_G' \Theta_{GB}) = \text{diag}(\mathbf{1}_A' \Theta_{AB})$, and therefore $\tilde{\Theta}_{GB} = \Theta_{GB} [\text{diag}(\mathbf{1}_A' \Theta_{AB})]^{-1}$.

Now,

$$\begin{aligned} \tilde{\Theta}_{AG} \tilde{\Theta}_{GB} &= \Theta_{AG} \tilde{\Theta}_{GB} \\ &= \Theta_{AG} \Theta_{GB} [\text{diag}(\mathbf{1}_A' \Theta_{AB})]^{-1} \\ &= \Theta_{AB} [\text{diag}(\mathbf{1}_A' \Theta_{AB})]^{-1} \\ &= \tilde{\Theta}_{AB}. \end{aligned} \quad (6.1)$$

Therefore, using this construction, the standardized link matrix $\tilde{\Theta}_{AB}$ from U^A to U^B can always be factorized into the two matrices $\tilde{\Theta}_{AG}$ and $\tilde{\Theta}_{GB}$.

6.2 Strong Optimality: Statement of the Problem

As mentioned before, the optimality problem that we consider here is to minimize the variance (3.2) with respect to the standardized link matrix $\tilde{\Theta}_{AB}$. Now, using the factorization presented in section 6.1, we have

$$\begin{aligned}\text{Var}(\hat{Y}^B) &= \mathbf{Y}' \tilde{\Theta}'_{AB} \Delta_A \tilde{\Theta}_{AB} \mathbf{Y} \\ &= \mathbf{Y}' \tilde{\Theta}'_{GB} \tilde{\Theta}'_{AG} \Delta_A \tilde{\Theta}_{AG} \tilde{\Theta}_{GB} \mathbf{Y} \\ &= \mathbf{Y}' \tilde{\Theta}'_{GB} \Delta_G \tilde{\Theta}_{GB} \mathbf{Y}\end{aligned}\quad (6.2)$$

where $\Delta_G = \tilde{\Theta}'_{AG} \Delta_A \tilde{\Theta}_{AG}$.

For any standardized link matrix $\tilde{\Theta}_{AB}$, the factorization presented in section 6.1 always produces the same first factor $\tilde{\Theta}_{AG}$. Therefore, if we seek for some optimal matrix $\tilde{\Theta}_{AB, \text{opt}}$ that minimizes the variance (3.2), it is sufficient to optimize the second factor $\tilde{\Theta}_{GB}$. We would also like the optimal matrix $\tilde{\Theta}_{AB, \text{opt}}$ to produce unbiased estimates.

Let U_i^G be the subpopulation of U^G containing the N_i^G links to the unit i of U^B . Note that the subpopulations U_i^G are disjoint. Thus, without loss of generality, we can order the links from U^A to U^B so that, for every i , the links to unit i in U^B are indexed consecutively. Now, let $\tilde{\theta}_{GB,i}$ be the i^{th} column vector of the matrix $\tilde{\Theta}_{GB}$, $i = 1, \dots, N^B$. By construction, the vector $\tilde{\theta}_{GB,i}$ contains non null elements only for the N_i^G links to the unit i of U^B . Hence, letting $\tilde{\theta}_{GB,i}$ be a column vector of size N_i^G containing the non null elements of $\tilde{\theta}_{GB,i}$, we have

$$\tilde{\theta}_{GB,i} = \begin{bmatrix} \mathbf{0} \\ \tilde{\theta}_{GB,i} \\ \mathbf{0} \end{bmatrix}.$$

Similarly, let $\mathbf{1}_{G,i}$ be the column vector of size N^G containing 1's for N_i^G elements, and 0's elsewhere. Letting $\mathbf{1}_{G,i}$ be a column vector of size N_i^G containing 1's, we have

$$\mathbf{1}_{G,i} = \begin{bmatrix} \mathbf{0} \\ \mathbf{1}_{G,i} \\ \mathbf{0} \end{bmatrix}.$$

Now, for the GWSM from U^G to U^B to be unbiased, we need to have $\tilde{\theta}'_{GB,i} \mathbf{1}_{G,i} = 1$ for all i , or equivalently $\tilde{\theta}'_{GB,i} \mathbf{1}_{G,i} = 1$. All this together leads to the following optimization problem:

Find a matrix $\tilde{\Theta}_{GB, \text{opt}} = \{\tilde{\theta}_{GB, \text{opt}, 1}, \dots, \tilde{\theta}_{GB, \text{opt}, N^B}\}$ satisfying $\tilde{\theta}'_{GB, \text{opt}, i} \mathbf{1}_{G,i} = 1$ for all $i = 1, \dots, N^B$, and minimizing the quadratic form $\text{Var}(\hat{Y}^B) = \mathbf{Y}' \tilde{\Theta}'_{GB} \Delta_G \tilde{\Theta}_{GB} \mathbf{Y}$.

This problem turns out to be nothing else than the minimization of a positive quadratic form under linear constraints. This is a relatively standard and simple problem to solve. It is well known that a solution always exists and is unique if the form (6.2) is positive definite, or if the null subspace of $\tilde{\Theta}_{GB}$ is not included in the null-space of Δ_G .

The above optimization problem can be rewritten in a different form. Let $\Delta_{G, ii'}$ be the submatrix of Δ_G corresponding to the elements in positions g and g' if g has a link with unit i and g' has a link with unit i' . These matrices constitute a partition of Δ_G . Note that the matrices $\Delta_{G, ii}$ are symmetric, positive definite, and $\Delta'_{G, ii'} = \Delta_{G, i'i}$. With these notations, the optimization problem can be written as:

Minimize

$$\sum_{i=1}^{N^B} \sum_{i'=1}^{N^B} y_i y_{i'} \tilde{\theta}'_{GB, i} \Delta_{G, ii'} \tilde{\theta}_{GB, i'} \quad (6.3)$$

under the constraints $\tilde{\theta}'_{GB, i} \mathbf{1}_{G, i} = 1$ for all $i = 1, \dots, N^B$.

Minimization is achieved for vectors $\tilde{\theta}_{GB, \text{opt}, i}$ satisfying

$$y_i \sum_{i'=1}^{N^B} \Delta_{G, ii'} \tilde{\theta}_{GB, \text{opt}, i'} y_{i'} = \lambda_i \mathbf{1}_{G, i} \quad (6.4)$$

for all $i = 1, \dots, N^B$ and where λ_i are the Lagrange multipliers entering into the constrained minimization of (6.3). As we can see from (6.4), the optimal choice $\tilde{\theta}_{GB, \text{opt}, i}$ (and therefore $\tilde{\Theta}_{GB, \text{opt}}$) will depend in general explicitly on the vector \mathbf{Y} , which is not useful in practice. Observe that the set of λ_i depends also of the variable \mathbf{Y} . This will appear more explicitly in section 6.3. This is the reason why we will seek, instead of a strong optimization, for a weaker form of optimality that will lead to the existence of an "optimal" solution $\tilde{\Theta}_{GB, \text{opt}}$ (and $\tilde{\Theta}_{AB, \text{opt}}$) not depending on \mathbf{Y} .

6.3 Weak Optimality

Equations (6.4) must be valid for any vector \mathbf{Y} . In particular, a necessary condition is to hold for a particular variable of interest, such as $y_i = 1$ for a unit i of U^B and $y_{i'} = 0$ for all other units i' of U^B ($i' \neq i$). This leads to the necessary conditions (one for each of those particular variables) $\Delta_{G, ii} \tilde{\theta}_{GB, \text{opt}, i} = \lambda_i \mathbf{1}_{G, i}$. Assuming that $\Delta_{G, ii}$ is invertible, we then have $\tilde{\theta}_{GB, \text{opt}, i} = \lambda_i \Delta_{G, ii}^{-1} \mathbf{1}_{G, i}$. It can be shown that this is also a sufficient condition. Now, because $\tilde{\theta}'_{GB, \text{opt}, i} \mathbf{1}_{G, i} = 1$, we have $\lambda_i = 1 / \mathbf{1}'_{G, i} \Delta_{G, ii}^{-1} \mathbf{1}_{G, i}$. Therefore, a necessary and sufficient condition for equation (6.4) to be satisfied is when

$$\tilde{\theta}_{GB, \text{opt}, i} = \frac{\Delta_{G, ii}^{-1} \mathbf{1}_{G, i}}{\mathbf{1}'_{G, i} \Delta_{G, ii}^{-1} \mathbf{1}_{G, i}}. \quad (6.5)$$

This result corresponds to weak optimization in the following sense. The weight w_i given by (2.6) satisfies $E(w_i) = 1$ and moreover $E(w_i | i \in \Omega^B) = 1 / \pi_i^B$ where π_i^B is the inclusion probability of unit i in Ω^B , which is generally difficult or even impossible to compute in practice. Now, note that the Horvitz-Thompson estimator is characterized by $\text{Var}(w_i | i \in \Omega^B) = 0$. The weak optimization

obtained here consists in minimizing $\text{Var}(w_i | i \in \Omega^B)$ over all possible standardized link matrices $\tilde{\Theta}_{GB}$, or equivalently $\tilde{\Theta}_{AB}$. This variance is strictly positive for the cases where unit i of U^B is in position to receive more than a unique weight for different sample s^A . Moreover, using (6.3), the multiplier λ_i appears to be the variance of the weight w_i and is, therefore, always strictly positive (except, a case that we exclude, when unit i is selected with a weight equal to one).

6.4 Strong Optimality Independent of \mathbf{Y}

Weak optimality is a necessary condition for strong optimality independent of the vector \mathbf{Y} of a variable of interest. It provides the necessary form of the vectors $\tilde{\Theta}_{GB, \text{opt}, i}$ in (6.4). To get sufficient conditions for strong optimality independent of \mathbf{Y} , we go back to the equations (6.4). These equations need to be satisfied for all vectors \mathbf{Y} and they must therefore be satisfied for a particular variable of interest such as $y_i = 1$ for a unit i of U^B , $y_{i'} = 1$ for another unit i' of U^B , and $y_{i''} = 0$ for all other units i'' of U^B ($i'' \neq i' \neq i$). In that case, to satisfy equations (6.4), it is necessary to have the following relations for any i and i' :

$$\Delta_{G, ii'} \tilde{\Theta}_{GB, \text{opt}, i} + \Delta_{G, ii'} \tilde{\Theta}_{GB, \text{opt}, i'} = \lambda_{i'} \mathbf{1}_{G, i} \quad (6.6)$$

$$\Delta_{G, i'i'} \tilde{\Theta}_{GB, \text{opt}, i} + \Delta_{G, i'i'} \tilde{\Theta}_{GB, \text{opt}, i'} = \lambda_{i'} \mathbf{1}_{G, i'}.$$

As we must necessarily have weak optimality, we have $\Delta_{G, ii'} \tilde{\Theta}_{GB, \text{opt}, i} = \lambda_i \mathbf{1}_{G, i'}$. Considering the first line of (6.6), we then get

$$\begin{aligned} \Delta_{G, ii'} \tilde{\Theta}_{GB, \text{opt}, i'} &= (\lambda_{i'} - \lambda_i) \mathbf{1}_{G, i} \\ &= \Phi_{ii'} \mathbf{1}_{G, i}. \end{aligned} \quad (6.7)$$

Multiplying both sides of (6.7) by $\tilde{\Theta}'_{GB, \text{opt}, i}$, we obtain

$$\begin{aligned} \tilde{\Theta}'_{GB, \text{opt}, i} \Delta_{G, ii'} \tilde{\Theta}_{GB, \text{opt}, i'} &= \Phi_{ii'} \tilde{\Theta}'_{GB, \text{opt}, i} \mathbf{1}_{G, i} \\ &= \Phi_{ii'} \end{aligned}$$

since $\tilde{\Theta}'_{GB, \text{opt}, i} \mathbf{1}_{G, i} = 1$. Let Φ be the matrix with elements $\Phi_{ii'}$ off the diagonal and $\Phi_{ii} = \lambda_i$ on the diagonal. Using again (6.2), it can be shown that the optimal variance (whenever it exists) has the expression $\mathbf{Y}'\Phi\mathbf{Y}$.

Let us show that this set of conditions is also sufficient. Assume that (6.7) holds. Note that for $i = i'$, condition (6.7) is nothing else than (6.5) which gives the necessary values for the $\tilde{\Theta}_{GB, \text{opt}, i'}$. It is now straightforward to verify that (6.4) holds whatever the value of \mathbf{Y} and that we have obtained the strong optimality. Now, the values of λ_i depend on \mathbf{Y} , as well as the variance $\text{Var}(\hat{Y}^B)$, but we have that equations (6.4) always have the same solution (6.5) that

does not depend on \mathbf{Y} . We therefore have the following result:

Result 7:

The conditions $\Delta_{G, ii'} \tilde{\Theta}_{GB, \text{opt}, i'} = \Phi_{ii'} \mathbf{1}_{G, i}$ are necessary and sufficient for the existence of a standardized link matrix $\tilde{\Theta}_{GB, \text{opt}}$, or equivalently $\tilde{\Theta}_{AB, \text{opt}}$, that achieves strong optimality independent of the vector \mathbf{Y} of the variable of interest. The values in the columns of this strong optimal matrix are given by (6.5), which are the vectors $\tilde{\Theta}_{GB, \text{opt}, i}$ obtained from weak optimality.

It should be noted that since $\Delta_{G, ii} \tilde{\Theta}_{GB, \text{opt}, i} = \lambda_i \mathbf{1}_{G, i}$ (6.7) can be written in an equivalent way as

$$\Phi_{ii'}^* \tilde{\Theta}_{GB, \text{opt}, i} = \Delta_{G, ii}^{-1} \Delta_{G, ii'} \tilde{\Theta}_{GB, \text{opt}, i'} \quad (6.8a)$$

or

$$\Phi_{ii'}^* \mathbf{1}_{G, i} = \Delta_{G, ii} \Delta_{G, i'i'}^{-1} \mathbf{1}_{G, i'} \quad (6.8b)$$

where $\Phi_{ii'}^* = (\tilde{\Theta}'_{GB, \text{opt}, i} \Delta_{G, ii'} \tilde{\Theta}_{GB, \text{opt}, i'}) (\mathbf{1}'_{G, i} \Delta_{G, ii}^{-1} \mathbf{1}_{G, i'})$ and $\Phi_{ii'}^* = (\tilde{\Theta}'_{GB, \text{opt}, i} \Delta_{G, ii} \tilde{\Theta}_{GB, \text{opt}, i'}) (\mathbf{1}'_{G, i'} \Delta_{G, i'i'}^{-1} \mathbf{1}_{G, i'})$. In some situations, these can proved to be easier to use than the expression (6.7) stated in Result 7.

6.5 Two Examples

We now present two examples that illustrate the preceding theory on weak optimality and strong optimality independent of \mathbf{Y} .

Example 1: Poisson Sampling

Let us suppose that the sample s^A is selected using Bernoulli or Poisson Sampling. In that case, the $N^A \times N^A$ matrix Δ_A is given by $\Delta_A = \text{diag}(1/\pi_j^A - 1)$. Considering the factorization of section 6.1, we have $\Delta_G = \tilde{\Theta}'_{AG} \Delta_A \tilde{\Theta}_{AG} = \tilde{\Theta}'_{AG} [\text{diag}(1/\pi_j^A - 1)] \tilde{\Theta}_{AG} = [\text{diag}((1/\pi_j^A - 1) \mathbf{1}_{A, jj})]$ where $\mathbf{1}_{A, jj}$ is a square matrix of size N_j^A , with N_j^A being the number of links (or “arrows”) starting from unit j of U^A . From Δ_G , we extract the submatrices $\Delta_{G, ii}$ that are, in the present case, diagonal. Each submatrix $\Delta_{G, ii}$ is given by $\Delta_{G, ii} = \text{diag}(1/\pi_g^A - 1)$, which is of size N_i^G . Note that each value $(1/\pi_g^A - 1)$ simply corresponds to a unit j of U^A that has previously been linked to the unit g of U^G , which is in turn linked to the unit i of U^B . Now, from (6.5), we directly obtain the optimal values $\tilde{\Theta}_{GB, \text{opt}, i}$ that minimize $\text{Var}(\hat{Y}^B)$, in the weak sense. These values are given by the vectors

$$\tilde{\Theta}'_{GB, \text{opt}, i} = \left\{ \frac{\pi_i^A}{(1 - \pi_i^A) \tau_i^G}, \dots, \frac{\pi_{N_i^G}^A}{(1 - \pi_{N_i^G}^A) \tau_i^G} \right\}$$

where

$$\tau_i^G = \sum_{g=1}^{N_i^G} \pi_g^A / (1 - \pi_g^A), i = 1, \dots, N^B.$$

The $\tilde{\theta}'_{GB, \text{opt}, i}$ are used to construct the vectors $\tilde{\theta}'_{GB, \text{opt}, i}$, and then the matrix $\tilde{\Theta}_{GB, \text{opt}} = \{\tilde{\theta}'_{GB, \text{opt}, 1}, \dots, \tilde{\theta}'_{GB, \text{opt}, N^B}\}$. Finally, after computing the optimal matrix $\tilde{\Theta}_{AB, \text{opt}} = \Theta_{AG} \tilde{\Theta}_{GB, \text{opt}}$, we obtain the optimal weights \mathbf{W}_{opt} using (2.6).

It should be noted that if the inclusion probabilities π_j^A are equal, we get

$$\tilde{\theta}'_{GB, \text{opt}, i} = \left\{ \frac{1}{N_i^G}, \dots, \frac{1}{N_i^G} \right\} = \frac{1}{N_i^G} \mathbf{1}_{GB, i},$$

where N_i^G is nothing else than the number of units of U^A linked to unit i of U^B . In other words, in the context of Bernoulli Sampling (i.e., Poisson Sampling with equal probabilities), to minimize the variance $\text{Var}(\hat{Y}^B)$, the choice of the values $\theta_{\text{opt}, ji}^{AB}$ should be given by 1 if there is a link between unit j of U^A and i of U^B , and 0 otherwise. This corresponds to the results obtained by Kalton and Brick (1995), Lavallée (2002), and Lavallée and Caron (2001).

Using Result 7, we now verify if conditions (6.7), (6.8a) or (6.8b) are satisfied for the optimal matrix $\tilde{\Theta}_{AB, \text{opt}}$ that we obtained through weak optimization. If it is the case, this matrix also provides strong optimality independent of the variable of interest y_i . First, we have

$$\Delta_{G, ii}^{-1} = \text{diag} \left(\frac{\pi_g^A}{1 - \pi_g^A} \right).$$

Also, each submatrix $\Delta_{G, ii'}$ of size $N_i^G \times N_{i'}^G$ has somewhat a diagonal structure, but "padded" with zeros. That is, a typical element of $\Delta_{G, ii'}$ is given by $(1/\pi_g^A - 1)$ on a part of the diagonal if both i and i' are linked to the same unit j of U^A (that is linked to unit g of U^G coming from the same j of U^A), and 0 otherwise. Because of this, if two units i and i' are not linked to the same units of U^A , then $\Delta_{G, ii'}$ is a matrix of zeros, and then the conditions (6.7), (6.8a) and (6.8b) are automatically satisfied. Referring to Figure 1, children $i = 2$ and $i' = 3$ of U^B are not related to the same parents j of U^A . If the selection of the parents is done using Poisson or Bernoulli Sampling, the 2×2 matrix $\Delta_{G, 23}$ will then contain only zeros, i.e.,

$$\Delta_{G, 23} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Because if this, the relations (6.7), (6.8a) or (6.8b) will be satisfied with $\Phi_{23} = 0$, expressing the fact that the weights of i and i' are not correlated.

If two units i and i' are linked to the same unit j of U^A , then, using (6.7), the column vector $\Delta_{G, ii'} \tilde{\theta}_{GB, \text{opt}, i'}$ contains the scalar $(\tau_{i'}^G)^{-1} = [\sum_{g=1}^{N_{i'}^G} \pi_g^A / (1 - \pi_g^A)]^{-1}$ for its first

$N_{i'}^B$ components, and 0 for the remaining $N_{i'}^B - N_{i'}^B$ ones (assuming $N_{i'}^B \geq N_{i'}^B$). Because the quantity $\Delta_{G, ii'} \tilde{\theta}_{GB, \text{opt}, i'}$ must be equal to $\Phi_{ii'} \mathbf{1}_{G, i}$ to satisfy (6.7), it must contain only the value $\Phi_{ii'}$. Since $\Phi_{ii'} = \tilde{\theta}'_{GB, \text{opt}, i'} \Delta_{G, ii'} \tilde{\theta}_{GB, \text{opt}, i'}$, this will occur only if the vector $\tilde{\theta}_{GB, \text{opt}, i} = [1]$, which means that there is only one link to unit i of U^B . As we can see, this is not a condition that will be satisfied in general and therefore, it can be said that in the case of Poisson Sampling, strong optimality independent from \mathbf{Y} will not occur in general.

As a conclusion, we might say that with Poisson or Bernoulli Sampling, the conditions (6.7), (6.8a) or (6.8b) will be satisfied in practice only when the units of U^A are linked to a single unit of U^B , as in the case of sampling households using a frame of individuals. In the other cases, the optimal matrix $\tilde{\Theta}_{AB, \text{opt}}$ obtained through weak optimality will not likely lead to strong optimization independent of \mathbf{Y} .

Example 2: Simple Random Sampling

Let us suppose that the sample s^A is selected using Simple Random Sampling. In that case, the $N^A \times N^A$ matrix Δ_A is given by

$$\Delta_A = \frac{N^A}{n^A} \frac{(N^A - n^A)}{(N^A - 1)} \left[\mathbf{I}_A - \frac{\mathbf{1}_A \mathbf{1}_A'}{N^A} \right].$$

Considering the factorization of section 6.1, we have

$$\begin{aligned} \Delta_G &= \tilde{\Theta}'_{AG} \Delta_A \tilde{\Theta}_{AG} \\ &= \frac{N^A}{n^A} \frac{(N^A - n^A)}{(N^A - 1)} \times \tilde{\Theta}'_{AG} \left[\mathbf{I}_A - \frac{\mathbf{1}_A \mathbf{1}_A'}{N^A} \right] \tilde{\Theta}_{AG} \\ &= \frac{N^A}{n^A} \frac{(N^A - n^A)}{(N^A - 1)} \times \left[\text{diag}(\mathbf{1}_{A, jj}) - \frac{\mathbf{1}_G \mathbf{1}_G'}{N^A} \right] \end{aligned} \quad (6.9)$$

where $\mathbf{1}_{A, jj}$ is a square matrix of size N_j^A , with N_j^A being the number of links (or "arrows") starting from unit j of U^A . From Δ_G , we extract the submatrices $\Delta_{G, ii'}$. Each submatrix $\Delta_{G, ii}$ is given by

$$\Delta_{G, ii} = \frac{N^A}{n^A} \frac{(N^A - n^A)}{(N^A - 1)} \times \left[\mathbf{I}_{G, i} - \frac{\mathbf{1}_{G, i} \mathbf{1}_{G, i}'}{N^A} \right],$$

which is of size N_i^G . Then, using a matrix result that can be found, amongst others, in Jazwinski (1970), we get

$$\Delta_{G, ii}^{-1} = \frac{(N^A - 1)}{(N^A - n^A)} \frac{n^A}{N^A} \times \left[\mathbf{I}_{G, i} + \frac{1}{(N^A - N_i^G)} \mathbf{1}_{G, i} \mathbf{1}_{G, i}' \right].$$

Now, from (6.5), we directly obtain the optimal values

$$\tilde{\theta}_{GB, \text{opt}, i} = \frac{1}{N_i^G} \mathbf{1}_{G, i}$$

that minimize $\text{Var}(\hat{Y}^B)$, in the weak sense, $i = 1, \dots, N^B$. These values are used to construct the vectors $\hat{\theta}'_{GB, \text{opt}, i^B}$, and then the matrix $\hat{\Theta}_{GB, \text{opt}} = \{\hat{\theta}_{GB, \text{opt}, 1^B}, \dots, \hat{\theta}_{GB, \text{opt}, N^B}^B\}$. Finally, after computing the optimal matrix $\hat{\Theta}_{AB, \text{opt}} = \Theta_{AG} \hat{\Theta}_{GB, \text{opt}}$, we obtain the optimal weights \mathbf{W}_{opt} using (2.6).

Again, this result is an important one because it goes directly in the direction of the results of Kalton and Brick (1995), Lavallée (2002), and Lavallée and Caron (2001). That is, with Simple Random Sampling, the optimal choice of $\theta_{\text{opt}, ji}^{AB}$ should be 1 if there is a link between unit j of U^A and i of U^B , and 0 otherwise.

Using Result 7, we now verify if the conditions (6.7), (6.8a) or (6.8b) for strong optimality independent of y_i are satisfied for the optimal matrix $\hat{\Theta}_{AB, \text{opt}}$ that we obtain through weak optimization. First, each submatrix $\Delta_{G, ii'}$ of size $N_i^G \times N_{i'}^G$ is given by

$$\Delta_{G, ii'} = \frac{N^A}{n^A} \frac{(N^A - n^A)}{(N^A - 1)} \times \left[\mathbf{H}_{G, ii'} - \frac{\mathbf{1}_{G, i} \mathbf{1}_{G, i'}'}{N^A} \right]$$

where $\mathbf{H}_{G, ii'}$ is a $N_i^G \times N_{i'}^G$ diagonal matrix of ones, "padded" with zeros. Exactly on the same pattern as in example 1, a typical element of $\mathbf{H}_{G, ii'}$ is given by 1 if both i and i' are linked to the same unit j of U^A (that is linked to unit g of U^G), and 0 otherwise. Therefore, we can easily see in which cases the conditions (6.7), (6.8a) or (6.8b) can be satisfied. In fact, because all components of $\hat{\theta}_{GB, \text{opt}, i}$ are equal, $\Delta_{G, ii'} \hat{\theta}_{GB, \text{opt}, i'}$ is a vector proportional to the sum of the lines of $\Delta_{G, ii'}$, i.e., the sum of the lines of

$$\left[\mathbf{H}_{G, ii'} - \frac{\mathbf{1}_{G, i} \mathbf{1}_{G, i'}'}{N^A} \right].$$

But (6.7) says that this vector must have the same components. This is possible if and only if the matrix $\mathbf{H}_{G, ii'}$ contains only zeros, or if it is of dimension 1×1 , which occurs when both i and i' are each linked to only one element of U^A . Therefore, as for Poisson Sampling, strong optimality independent of \mathbf{Y} does not occur in general for Simple Random Sampling.

7. Conclusion

In the present paper, we discussed the use of Indirect Sampling together with the method developed to obtain estimation weights: the Generalized Weight Share Method (GWSM). We then showed the following properties of the GWSM: unbiasedness, the variance computation and transitivity. We presented after a section on the use of the GWSM when the links between the populations U^A and U^B are expressed by ones and zeros, i.e., there is a link or

there is not. The section after was devoted to results that are obtained with different forms of link matrices. Finally, we assessed the problem of optimality, i.e., the choice of optimal values to express the links between U^A and U^B in order to minimize the variance of the estimates issued from the GWSM. We have distinguished two kind of optimization: weak and strong optimization.

Weak optimization consists in finding the values of the links to be used in order to minimize, for each unit, the variance of the weights provided by the GWSM. The solution is always uniquely defined, easy to compute and to implement in practice. Weak optimization is also a necessary condition for strong optimization. Strong optimization consists in finding the values of the links in order to minimize the variance of estimation for the total of any variable of interest y . It does not exist for all sampling designs and type of links between the populations U^A and U^B . It also depends on somewhat complicated relations.

We recommend the use of weak optimization because of its flows naturally and the fact that it is very easy to use. Moreover, if our estimation problem can be as well optimized in the strong sense, we will have achieved it through weak optimization, even if it was not demonstrated!

Acknowledgements

The authors would like to thank all the people that showed an interest in Indirect Sampling, and especially in the GWSM. They motivated the writing of this paper that goes beyond what was made previously on this subject.

References

- Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons, Inc. 135-159.
- Déville, J.C., (1998). Comment attraper une population en se servant d'une autre. *Insee méthodes*, n°84-85-86, *Actes des Journées de méthodologie statistique des 17-18 mars 1998*, 63-82.
- Jazminski, A.H. (1970). *Stochastic Processes and Filtering Theory*. New York: Academic Press.
- Kalton, G., and Brick, J.M. (1995). Weighting Schemes for Household Panel Surveys. *Survey Methodology*, 21, 1, 33-44.
- Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 1, 25-32.
- Lavallée, P. (2002). *Le Sondage Indirect, ou la Méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles, Brussels.

- Lavallée, P., and Caron, P. (2001). Estimation using the generalised weight share method: The case of record linkage. *Survey Methodology*, 27, 2, 155-169.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Searle, S.R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.

Extension of the Indirect Sampling Method and its Application to Tourism

Jean-Claude Deville and Myriam Maumy-Bertrand¹

Abstract

A survey of tourist visits originating intra and extra-region in Brittany was needed. For concrete material reasons, "border surveys" could no longer be used. The major problem is the lack of a sampling frame that allows for direct contact with tourists. This problem was addressed by applying the *indirect sampling method*, the weighting for which is obtained using the *generalized weight share method* developed recently by Lavallée (1995), Lavallée (2002), Deville (1999) and also presented recently in Lavallée and Caron (2001). This article shows how to adapt the method to the survey. A number of extensions are required. One of the extensions, designed to estimate the total of a population from which a Bernoulli sample has been taken, will be developed.

Key Words: Generalized weight share method; Incomplete frame and multiple frames.

1. Introduction

A "border survey" of extra-region tourist visits in Brittany (those not by residents of Brittany) was conducted over the period from April to September 1997. The Observatoire Régional du Tourisme de Bretagne and the Comités Départementaux de Tourisme were interested in doing another one. Unfortunately, they no longer had the means to gather a certain mass of data at the regional or intra-regional borders because the police forces were no longer interested in collaborating on roadside surveys.

For this reason, the Observatoire Régional du Tourisme de Bretagne, with the assistance of a technical committee comprised of methodologists and field operators, decided to introduce a new survey methodology to replace the "border survey" methodology. In addition, evaluation of intra-regional tourism (of residents of Brittany vacationing in Brittany, for example) is vital to identifying development factors.

One of the major problems is the lack of a sampling frame that allows direct communication with tourists. This problem was addressed by using an approach previously used in the Asturias in Spain (Valdés, De La Ballina, Aza, Loredó, Torres, Estébanez, Domínguez and Del Valle (2001) and Torres Manzanera, Sustacha Melijosa, Menéndez Estébanez and Valdés Pelaáez (2002)), which involves sampling services intended mainly for tourists and asking them questions at the various locations of these many tourist service sites. Obviously, a tourist may use one or more of the services in the sampling frame once or several times during the survey period in question. To be able to estimate the parameters of interest with respect to tourists, it must be possible to conduct a rigorous sample of certain services and then link the set of weights of the sampled

services to the set of weights of the tourists who used these services. The purpose of this article is to present a method that makes this calculation possible. This method relies mainly on the generalized weight share method (GWSM) developed by Lavallée (1995), Lavallée (2002) and Deville (1999).

2. Generalized Weight Share Method

We will briefly review the principle of the *generalized weight share method* (GWSM). For more information, see Lavallée (1995), Lavallée (2002) and Deville (1999).

We will let U^A be a finite population containing N^A units, where each unit is denoted by j and U^B is a finite population containing N^B units, where each unit is denoted by i . The correspondence between U^A and U^B can be represented by a matrix of links $\Theta_{AB} = [\theta_{ji}^{AB}]$, of size $N^A \times N^B$ where each element $\theta_{ji}^{AB} \geq 0$. In other words, the unit j of U^A is linked to unit i of U^B provided that $\theta_{ji}^{AB} > 0$; otherwise, there is no link between these two units.

In the case of the indirect survey, we select the sample s^A of n^A units from U^A based on a given sampling design. Let $\pi_j^A > 0$, be the probability of selection of the unit j . For each unit j selected in s^A , we identify the units i of U^B for which $\theta_{ji}^{AB} > 0$. Then we let s^B , be all of the n^B units of U^B identified using the units $j \in s^A$, that is,

$$s^B = \{i \in U^B; \exists j \in s^A \text{ and } \theta_{ji}^{AB} > 0\}.$$

For each unit i of s^B , a variable of interest y_i is measured.

It is assumed that, for any unit j of s^A , it is possible to obtain the values of θ_{ji}^{AB} for $i=1, \dots, N^B$ by a direct interview or from an administrative source. For any unit i

1. Jean-Claude Deville, Laboratoire de Statistique d'Enquêtes, ENSAI/CREST, Campus de Ker-Lann, 35170 BRUZ (France). E-mail: deville@ensai.fr;
Myriam Maumy-Bertrand, Laboratoire de Statistique, Université Louis Pasteur, 7, rue René Descartes 67084 STRASBOURG Cedex (France). E-mail: mmaumy@math.u-strasbg.fr.

identified of U^B (or only of s^B), it is assumed that we can obtain the values of θ_{ji}^{AB} for $j = 1, \dots, N^A$. For this reason, it is not necessary to know the values of θ_{ji}^{AB} for all of the matrix of links Θ_{AB} . Indeed, we only need to know the values of θ_{ji}^{AB} for lines j of Θ_{AB} , where $j \in s^A$, and for columns i of Θ_{AB} where $i \in s^B$.

For example, if the purpose is to estimate a variable of interest Y^B of target population U^B , where

$$Y^B = \sum_{i=1}^{N^B} y_i, \quad (2.1)$$

with y_i measured according to the aggregate U^B . We then use an estimator in the form

$$\hat{Y}^B = \sum_{i=1}^{N^B} w_i y_i, \quad (2.1)$$

where w_i is the estimated weight of unit i of s^B , with $w_i = 0$ for $i \notin s^B$. To obtain an unbiased estimate of a variable of interest Y^B , we must use as weight w_i the inverse of the probability of selection π_i^B of unit i . As mentioned in Lavallée (1995) and Lavallée (2002), it is generally difficult, if not impossible, to obtain these probabilities. Consequently, we turn to the GWSM, where the weights are given by

$$w_i = \sum_{j \in s^A} \frac{\tilde{\theta}_{ji}^{AB}}{\pi_j^A},$$

where $\tilde{\theta}_{ji}^{AB} = \theta_{ji}^{AB} / \sum_{j=1}^{N^A} \theta_{ji}^{AB}$. Using this construction, the estimator \hat{Y}^B is unbiased. Similarly, it is possible to calculate and estimate the variance of this estimator because it is the same as that of

$$\sum_{j \in s^A} \frac{z_j}{\pi_j^A},$$

with $z_j = \sum_{i=1}^{N^B} \tilde{\theta}_{ji}^{AB} y_i$.

3. Tourism Survey in an Open Environment

3.1 Survey Objectives

The principle of the survey is as follows:

"reach tourists (foreigners or French citizens whether or not they live in Brittany) through services aimed at meeting the basic or specific needs"

such as accommodation, food, leisure activities and transportation.

3.2 Population of Interest

We will let G be a *geographic field* (the four provinces of Brittany) and P be a *reference period* (in this case, it is from February 2005 to December 2005).

A *tourist* is defined as a person who spent at least one night in G outside his principle residence (tourist-night).

For a tourist, a *trip* is an period *sej* of P , the length of the cardinal of *sej* noted as $|sej|$, during which the tourist spends all his nights in G outside his principle residence, the nights immediately before or after the trip *sej* having been spent outside G (or at the principle residence.).

A *tour* is a group of tourists (tourist household) sharing the same trip and with the same accommodation during the trip. The term tourist household will also be used through a slight misuse of the terminology (the same tourist household can have several tours over a period, but we have no way to distinguish them).

The *statistical unit* i of the survey is the tour.

The *sub-units of the survey* are the trips, tourists and tourist-nights. A tour i consists of n_i tourists during a trip of duration $|sej|$ and thus $n_i \times |sej|$ tourist-nights. Here population U^B is therefore the aggregate of the tours in G during P . ($sej \cap P \neq \emptyset$).

3.3 Survey Sampling Design

To use the GWSM, the theoretical population U^A is formed by a "services" aggregate. In this survey, these services consist of:

- Purchases in bakeries, being the first stratum of U^A .
- Visits to a set of well known cultural, recreational or family sites. In practice, for each of them, a "mandatory pass point" has been defined. It consists of the total number of people passing by this point, which is the second stratum of U^A .
- The number of people exiting Brittany by way of the La Gravelle highway toll, which accounts for 80% of the exits by tourists from Brittany by car. This method of transport itself accounts for 80% of the trips by non-resident of Brittany. People passing this point constitute the third stratum of U^A .

In other words, the *sampling frame* is formally constructed of three strata:

1. purchases in bakeries;
2. visits to a set of sites typical of Brittany;
3. people at the La Gravelle highway toll.

In the *first stratum*, we use a three-stage sample:

- a sample of bakeries;
- a sample of survey days;
- a sample of clients in the bakery on a given day.

In the *second stratum*, we use a two-stage sample:

- a sample of survey days;
- a sample of people who pass through one of the 16 chosen sites on a given day.

Lastly, in the *third stratum*, we use a two-stage sample:

- a sample of survey days;
- a sample of people who pass through the La Gravelle highway toll on a given day.

It is acknowledged that any tourist household consumes at least one of the “services” (bakery purchases, visits to typical Brittany sites, the La Gravelle highway toll), or at least, that very few households do not consume any of them.

Each sampling (bakery, days, “service”) requires specific techniques and it would take considerable time to provide details on each of them. Nevertheless, we will provide the following key technical elements:

- bakeries are sampled using a traditional design stratified geographically (five strata: coastal area of four Brittany departments, the interior of Brittany). In each stratum, the bakeries are sampled with probabilities proportional to their “tourist potential” constructed from their business revenue, the tourist accommodation capacity, and the number of principal residences in the commune to which they belong. This was the theoretical approach, but in practice, the sample was somewhat “forced” by unforeseen circumstances (refusal of bakers, closures during certain period, for example).
- The sites are not sampled, but rather selected for their notoriety and the technical possibility of identifying a “mandatory pass point” (sometimes approximate).
- For each bakery, each site and the La Gravelle highway toll, we defined completely homogeneous “clusters of days” in each period P . A cluster was assigned randomly to each bakery, site and the La Gravelle highway toll. In practice, this means that a full-time enumerator is mobilized for several clusters.
- For each “service”, tourists are sampled using the normal techniques of random selection of arrivals: pseudo-systematic sample because, while the enumerator is handing out one questionnaire, other people are going by without being counted. This means that the total number of visitors cannot be estimated directly. If a site is accessible through a ticket booth (museum or chateau, for example), the sampling relies on this means. Ultimately, the sample of users of a “service” on a given day is considered a Bernouilli sample, that is, a simple random sample if we know the size of the population (the number of visitors on a given day).

Comments 3.1. The definition of *tourist* itself is linked to accommodation and it seems natural to use a frame directly

related to this service. Practice shows that this is difficult to achieve.

To begin with, there is no correct sampling frame for non-commercial accommodation (relatives, friends, secondary residence) or for seasonal furnished rentals.

In the case of hotels, campgrounds and family holiday homes, the trials runs in summer 2004 revealed the existence of catastrophic bias due to the intervention of hotel owners in the survey selection process. The hoteliers did not respect the random sample instructions and “essentially” distributed the questionnaires to their best clients. This part of the survey had to be set aside and replaced by the count through the La Gravelle highway toll, which is regularly subject to honest quality surveys by various organizations.

The questionnaires collected at the bakeries and at the Brittany tourism sites during summer 2004 apparently produced good qualitative and quantitative results regarding the various modes of accommodation.

Food consumption would undoubtedly have been captured better by questionnaires at the exit of supermarkets, but the problem there lies in the heterogeneity of these establishments and in the cutthroat competition between them; group C ... agrees to the surveys in its establishments only if group I ... is excluded! In contrast, the collaboration of local bakers in the survey was excellent.

Comments 3.2. By the very definition of the method used, we operate formally within the context of sampling from multiple frames. The problem has given rise to considerable literature (Hartley (1962), Lund (1968) and Hartley (1974) for a start). The GWSM applies to this problem by simply considering each sampling frame as a stratum provided that it is possible to identify for each unit sampled all of frames of which it is a part. This approach provides a rigorous and unique design-based solution to this problem. This comment is worthy of its own article, but the authors know that it is not worth the trouble: an idea that can be explained in ten lines does not need an article or a book for it to survive.

4. Parameters of Interest

Application F , which links to any service j during the reference period P in the three types of establishments of the survey coverage tour i that used this service, is defined as:

$$\begin{array}{ll} F : \text{services} & \rightarrow \quad \text{tour} \\ j & \rightarrow F(j) = i. \end{array}$$

We will let U^B , be the population of tours i of reference period P . This population of interest U^B is the image by F of the aggregate of services during reference period P

in the three types of establishments of the survey coverage. Population U^A is the image by F^{-1} of the aggregate of tours during reference period P . For all $i \in U^B$, we define $R_i(B) = \text{card}(F^{-1}(i))$, the number of antecedents of i during the survey period, that is, the number of services j used by the given tourist household i .

The *parameters of interest* can be totals, sizes or ratios. Let us assume, for example, that we are interested in the estimate of a total relative to a variable y defined on population U^B ,

$$Y^B = \sum_{i \in U^B} y_i. \quad (4.1)$$

A specific example of these totals is the size of U^B , written N^B and defined by

$$N^B = \text{card}(U^B) = \sum_{i \in U^B} 1.$$

For example, Y^B can be the number of people who practiced this activity, the total budget spent by the tourist household in Brittany, the geographic origin of the tourist households, or the number of days that the tourist household spends in Brittany. It should be noted that for many variables, the total Y^B depends on the size of the tourist household, that is, the number of people who make up this group and on the length of the trip (only those days spent in Brittany).

Now, we can write

$$Y^B = \sum_{i \in U^B} y_i = \sum_{l=1}^3 \sum_{a_l \in A_l} \sum_{d_l \in D_l} \sum_{j \in C_{d_l}} z_j, \quad (4.2)$$

where

$$z_j = \frac{y_i}{R_i(B)}, \text{ for } j \in F^{-1}(i),$$

where

- A_1 : the aggregate of bakeries in the survey coverage identified by index a_1
- A_2 : the 16 visit locations in the survey coverage identified by index a_2
- A_3 : the La Gravelle highway toll identified by index a_3
- D_l : the aggregate of survey days, identified by index d_l in an establishment a_l of A_l , for the variant of 1 to 3
- C_{d_l} : the aggregate of services in an establishment a_l of A_l of day d_l of D_l identified by index j .

5. Unbiased Estimates of a Total

In the previous paragraph, we showed that the total of interest is written as a total over the aggregate of the services in the coverage. Let us assume that we have a sample of respondent services j , to which we can link

sampling weight δ_j . These weights are assumed to be unbiased because the sample of services follows the canons of a multi-stage sample, each component sample being unbiased.

To make the notations easier to read, we will not show below all stages of the sample draw based on establishment a_l . Let:

- s^B : be the aggregate of tourist household i corresponding to the aggregate of services sampled during the survey period
- s_{A_l} : be the aggregate of sampled establishments
- s_{D_l} : be the aggregate of days sampled in establishment a_l
- s_{d_l} : be the sub-sample of services j corresponding to establishment day a_l .

Since we have a set of sampling weights δ_j for the respondent services, and if we know $R_i(B)$, we can estimate the unbiased total Y^B by

$$\hat{Y}^B = \sum_{i \in s^B} w_i y_i \quad (5.1)$$

where

$$w_i = \frac{\sum_{l=1}^3 \sum_{s_{A_l}} \sum_{s_{D_l}} \sum_{s_{d_l}} \delta_j}{R_i(B)}.$$

This gives us an estimate of the population of tourist households. This formula is none other than that given by the GWSM mentioned in section 2. Note that $U^A = U^{A_1} \cup U^{A_2} \cup U^{A_3} = \bigcup_{j=1}^3 U^{A_j}$, $\theta_{ji}^{AB} = 1$ of service j was used by tour i and then $\delta_j = 1/\pi_j^A$.

The variance can be estimated using the same principles (see Lavallée (2002)). We will not go into the details here because it is simply an application of general principles that requires somewhat onerous calculations.

Furthermore, using auxiliary information in the form of totals, whether in populations U^{A_l} or in population U^B , does not pose any particular problems for the point estimation or the estimation of the variance (see Lavallée (2002)).

Comments 5.1. The procedure we have just described for sharing weights may be considered naïve. In fact, we know how to optimize the links matrix Θ_{AB} as shown in Deville and Lavallée (2006). The application of the Brittany survey is described in Deville, Lavallée and Maumy (2005).

6. An example of a Specific Problem: Visit Points in Open Country

As has already been mentioned, developing the survey of tourism in Brittany required many complementary studies.

We have already mentioned the optimization of weight sharing. Using auxiliary data related to the various frames and to the various stages of the sampling is another task. In this section, we want to focus on estimating some of these auxiliary data, in particular for visits to tourism sites in open country.

In certain cases, we unfortunately do not know the total number of people, denoted as $T_p^{A_2}$, coming to the site on a given day. In effect, in aggregate A_2 , we do not know all the services (here the number of visits) of the population. It is therefore not possible to obtain $\pi_j^{A_2}$ directly and therefore δ_j for $j \in A_2$. To overcome this problem, we estimate the number of daily visitors in order to deduct $\hat{\pi}_j^{A_2} = n_{A_2} / \hat{T}_p^{A_2}$.

Our next step was to develop two approaches to estimating the number of daily visitors for sites accessible by vehicles only (or almost!). The first approach is based on a vehicle sampling system intended to estimate the number of visitors to the site. The second approach uses a sampling of visitors and is aimed at estimating the same quantity by interviewing individuals who give the number of people who travelled with him or her in the vehicle. These two approaches are developed in sections 7 and 8 below.

7. Constructing an Estimator of the Number of Visitors Using a Vehicle Sample

In this paragraph, we examine the approach where an enumerator counts the number of occupants in vehicles that break the line of an electronic eye, or an equivalent system has been set up to count vehicles for which the total number, written as T_v , is known with a virtually negligible measurement error.

7.1 Definition and Variance of $\hat{T}_p^{A_2}$

The total number of vehicles equals

$$T_v = \sum_{\kappa=1, \dots} t_{\kappa} = \sum_{l \in U_v} 1, \quad (7.1)$$

where t_{κ} represents the number of vehicles carrying κ persons and U_v the vehicle universe.

Comments 7.1. To make the notations easier to read, we will use here and until the end this article T_p to denote $T_p^{A_2}$.

The total number of people visiting the site equals

$$T_p = \sum_{\kappa=1, \dots} \kappa t_{\kappa} = \sum_{k \in U_p} 1, \quad (7.2)$$

where U_p denotes the universe of people. We also have the equation

$$T_p = \sum_{l \in U_v} v_l, \quad (7.3)$$

where v_l is the number of people in vehicle l .

As mentioned in the previous section, the total number of people T_p is unknown. Consequently, we must construct an estimator of T_p . If we let \hat{T}_p be π -estimator based on s_v , a simple random sample of vehicles of size n and with a probability of inclusion n/T_v

$$\hat{T}_p = \frac{T_v}{n} \sum_{l \in s_v} v_l = T_v \bar{v}, \quad (7.4)$$

assuming

$$\bar{v} = \frac{1}{n} \left(\sum_{l \in s_v} v_l \right).$$

It is clear that \hat{T}_p is an unbiased estimator of the total number of people T_p and that \bar{v} is an unbiased estimate of the average number \bar{V} of people in a vehicle.

The variance of \hat{T}_p is therefore equal to

$$\begin{aligned} \text{Var}[\hat{T}_p] &= T_v^2 \left(\frac{1}{n} - \frac{1}{T_v} \right) S_v^2 \\ &= \frac{1}{n} T_v^2 S_v^2 - T_v S_v^2, \end{aligned} \quad (7.5)$$

where S_v^2 denotes the corrected variance of population U_v .

7.2 Constructing an Estimator of a Variable of Interest in the Case of a Vehicle Sample

We want to estimate a variable of interest Y of population U_p written as

$$Y = \sum_{k \in U_p} y_k, \quad (7.6)$$

where y_k is the variable of interest measured in the final questionnaire. Let \hat{Y} be π -estimator defined by

$$\hat{Y} = \sum_{k \in s_p} w_k^p y_k, \quad (7.7)$$

where weight w_k^p is equal to \hat{T}_p/m . Consequently, estimator \hat{Y} can be written

$$\hat{Y} = \frac{\hat{T}_p}{m} \sum_{k \in s_p} y_k = \hat{T}_p \bar{y} \quad (7.8)$$

assuming

$$\bar{y} = \frac{1}{m} \left(\sum_{k \in s_p} y_k \right).$$

Subsequently, variables \hat{T}_p and \bar{y} will be assumed to be independent. The assumption is realistic, because we use two independent enumerators in the field.

7.2.1 Calculation of the Variance of the Estimator \hat{Y}

According to Huygens' theorem (1673), conditioning on sample s_v , we get

$$\begin{aligned}
V_Y &= \text{Var}[\hat{Y}] \\
&= \bar{Y}^2 \text{Var}[\hat{T}_p] + T_p^2 \text{Var}[\bar{Y}] \\
&\quad + \text{Var}[\hat{T}_p] \text{Var}[\bar{Y}].
\end{aligned} \quad (7.9)$$

In the present case, we liken the sample to a simple random sampling without replacement. Equation (7.9) thus becomes

$$\begin{aligned}
V_Y &= \bar{Y}^2 \left(\frac{1}{n} T_p^2 S_Y^2 - T_p S_Y^2 \right) \\
&\quad + T_p^2 \left(\frac{1}{m} S_Y^2 - \frac{S_Y^2}{T_p} \right) \\
&\quad + \left(\frac{1}{n} T_p^2 S_Y^2 - T_p S_Y^2 \right) \left(\frac{1}{m} S_Y^2 - \frac{S_Y^2}{T_p} \right),
\end{aligned}$$

with $S_Y^2 = 1 / (T_p - 1) \sum_{k \in U_p} (y_k - \bar{Y})^2$. Reorganizing the terms gives

$$\begin{aligned}
V_Y &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_p} \right) T_p^2 S_Y^2 \frac{1}{n} \\
&\quad + (T_p^2 - T_p S_Y^2) S_Y^2 \frac{1}{m} \\
&\quad + T_p^2 S_Y^2 S_Y^2 \frac{1}{nm} + \frac{T_p}{T_p} S_Y^2 S_Y^2 \\
&\quad - \bar{Y}^2 T_p S_Y^2 - T_p S_Y^2.
\end{aligned}$$

The next step is to determine the allocation of the sample sizes s_p and s_Y that minimizes the variance of estimator \hat{Y} for fixed population sizes T_p and T_Y .

We must therefore minimize equation (7.10) in n, m subject to

$$C_Y n + C_p m = C,$$

where C_Y denotes the cost (in time for example) of the questionnaires related to vehicles, C_p the cost (in time) of the questionnaires related to people, and C the total cost.

The Lagrangian equation can be written as

$$\begin{aligned}
L(n, m, \lambda) &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_p} \right) T_p^2 S_Y^2 \frac{1}{n} \\
&\quad + (T_p^2 - T_p S_Y^2) S_Y^2 \frac{1}{m} \\
&\quad + T_p^2 S_Y^2 S_Y^2 \frac{1}{nm} + \frac{T_p}{T_p} S_Y^2 S_Y^2 \\
&\quad - \bar{Y}^2 T_p S_Y^2 - T_p S_Y^2 \\
&\quad + \lambda (C_Y n + C_p m - C).
\end{aligned} \quad (7.11)$$

Taking the partial derivatives with respect to variables n, m, λ and setting them equal to zero gives

$$\begin{aligned}
\frac{\partial L}{\partial n}(n, m, \lambda) &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_p} \right) T_p^2 S_Y^2 \left(-\frac{1}{n^2} \right) \\
&\quad + T_p^2 S_Y^2 S_Y^2 \left(-\frac{1}{nm^2} \right) \\
&\quad + \lambda C_Y = 0,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial L}{\partial m}(n, m, \lambda) &= (T_p^2 - T_p S_Y^2) S_Y^2 \left(-\frac{1}{m^2} \right) \\
&\quad + T_p^2 S_Y^2 S_Y^2 \left(-\frac{1}{nm^2} \right) \\
&\quad + \lambda C_p = 0,
\end{aligned}$$

$$\frac{\partial L}{\partial \lambda}(n, m, \lambda) = C_Y n + C_p m - C = 0.$$

After calculations, we get a third-degree equation in n that is written

$$\begin{aligned}
\lambda C_Y^2 n^3 - \lambda C_Y C n^2 \\
- C_Y T_p^2 S_Y^2 \left(\bar{Y}^2 - \frac{S_Y^2}{T_p} \right) n \\
+ T_p^2 S_Y^2 \left(C \left(\bar{Y}^2 - \frac{S_Y^2}{T_p} \right) + C_p S_Y^2 \right) = 0.
\end{aligned}$$

This third-degree equation in n allows a real solution that can be determined using numeric methods.

Using the same reasoning, we get a third-degree equation in m

$$\begin{aligned}
\lambda C_p^2 m^3 - \lambda C_p C m^2 \\
- C_p S_Y^2 (T_p^2 - T_p S_Y^2) m \\
+ S_Y^2 (C(T_p^2 + T_p S_Y^2) + C_Y T_p^2 S_Y^2) = 0.
\end{aligned}$$

7.2.2 Simplified Case

To simplify the variance calculation of estimator \hat{Y} , we can make an approximation in equation (7.10). In effect, we can assume that term $1/nm$ is negligible before terms $1/n$ and $1/m$.

This then gives us the following transformation of equation (7.10)

$$\begin{aligned}
V_Y &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_p} \right) T_p^2 S_Y^2 \frac{1}{n} \\
&\quad + (T_p^2 - T_p S_Y^2) S_Y^2 \frac{1}{m} \\
&\quad + \frac{T_p}{T_p} S_Y^2 S_Y^2 - \bar{Y}^2 T_p S_Y^2 \\
&\quad - T_p S_Y^2.
\end{aligned} \quad (7.12)$$

The next step is determining the allocation of the sample sizes s_p and s_v that minimize the variance of estimator \hat{Y} for fixed population sizes T_p and T_v .

We must therefore minimize equation (7.12) in n, m subject to

$$C_v n + C_p m = C.$$

The Lagrangian equation can be written as

$$\begin{aligned} L(n, m, \lambda) = & \left(\bar{Y}^2 - \frac{S_Y^2}{T_p} \right) T_v^2 S_v^2 \frac{1}{n} \\ & + (T_p^2 - T_v S_v^2) S_v^2 \frac{1}{m} \\ & + \frac{T_v}{T_p} S_v^2 S_Y^2 - \bar{Y}^2 T_v S_v^2 \\ & - T_p S_Y^2 \\ & + \lambda (C_v n + C_p m - C). \end{aligned} \quad (7.13)$$

Taking the partial derivatives with respect to variables n, m, λ and setting them equal to zero gives

$$\begin{aligned} \frac{\partial L}{\partial n}(n, m, \lambda) &= \left(\bar{Y}^2 - \frac{S_Y^2}{T_p} \right) T_v^2 S_v^2 \left(-\frac{1}{n^2} \right) \\ &+ \lambda C_v = 0, \\ \frac{\partial L}{\partial m}(n, m, \lambda) &= (T_p^2 - T_v S_v^2) S_v^2 \left(-\frac{1}{m^2} \right) \\ &+ \lambda C_p = 0, \\ \frac{\partial L}{\partial \lambda}(n, m, \lambda) &= C_v n + C_p m - C = 0. \end{aligned}$$

After the calculations, we get

$$\begin{aligned} n_{\text{opt}} &= \frac{C}{\left(C_v + \sqrt{C_p C_v \frac{T_p S_Y^2 (T_p^2 - T_v S_v^2)}{T_v^2 S_v^2 (T_p \bar{Y}^2 - S_Y^2)}} \right)}, \\ m_{\text{opt}} &= \frac{C}{\left(C_p + \sqrt{C_p C_v \frac{T_v^2 S_v^2 (T_p \bar{Y}^2 - S_Y^2)}{T_p S_Y^2 (T_p^2 - T_v S_v^2)}} \right)}. \end{aligned}$$

8. Constructing an Estimator of the Number of Visitors Using a Sampling of Visitors

The previous method can be complicated and costly to use at certain sites. A simpler data collection method involves asking person k the number u_k of passengers in vehicle i that transported him or her. This number u_k is equal here to v_l for vehicle l that transported person k . This method has the further advantage of accurately capturing the number of passengers within the meaning of the survey (are babies counted?).

8.1 Definition of \hat{T}_p

Let us go back to the following equation

$$T_p = \sum_{l \in U_v} v_l,$$

where v_l denotes the number of passengers in vehicle l . Let us also recall

$$T_p = \sum_{l \in U_p} 1.$$

The average number of passengers in a vehicle \bar{V} can be expressed as

$$\bar{V} = \frac{\sum_{l \in U_v} v_l}{\sum_{l \in U_v} 1} = \frac{\sum_{\kappa=1, \dots} \kappa t_{\kappa}}{\sum_{\kappa=1, \dots} t_{\kappa}} = \frac{\sum_{\kappa=1, \dots} m_{\kappa}}{\sum_{\kappa=1, \dots} M_{\kappa} / \kappa}, \quad (8.1)$$

where t_{κ} is the number of κ -passenger vehicles and M_{κ} is the number of people who came in a κ -passenger vehicle.

We can use this last relation to obtain a new version of T_p

$$T_p = T_v \bar{V}. \quad (8.2)$$

Consequently, an estimator of T_p can be written as

$$\hat{T}_p = T_v \hat{\bar{V}}, \quad (8.3)$$

where the total number of vehicles T_v is perfectly known. Observing this expression, we see that, in order to know estimator \hat{T}_p , all that is required is to determine the quantity $\hat{\bar{V}}$. Let us therefore introduce the following estimator of \bar{V}

$$\hat{\bar{V}} = \frac{\sum_{\kappa \in s_p} m_{\kappa}}{\sum_{\kappa \in s_p} m\kappa / \kappa},$$

where m_{κ} is the number of people in the sample travelling in a κ passenger vehicle. Estimator $\hat{\bar{V}}$ can also be written as follows:

$$\hat{\bar{V}} = \frac{\sum_{k \in s_p} 1}{\sum_{k \in s_p} 1/u_k}$$

or as

$$\hat{\bar{V}} = \frac{m}{\sum_{k \in s_p} 1/u_k}. \quad (8.4)$$

The last equation makes it possible to write the following equation

$$\frac{1}{\hat{\bar{V}}} = \frac{1}{m} \sum_{k \in s_p} \frac{1}{u_k}. \quad (8.5)$$

This new quantity represents the empirical average of $1/u_k$ and $\hat{\bar{V}}$ is the harmonic average of u_k . It is also possible to calculate its variance, which is equal to

$$\text{Var}\left[\frac{1}{\hat{\bar{V}}}\right] = \left(\frac{1}{m} - \frac{1}{T_p}\right) S_{1/u}^2. \quad (8.6)$$

8.2 Calculating the Variance of Estimator \hat{T}_p Without a Vehicle Sample

Now we have to calculate the variance of estimator $\hat{\bar{V}}$ knowing (8.6). To this end, note that we can write

$$\begin{aligned} \frac{1}{\hat{\bar{V}}} &= \frac{1}{\bar{V} \left(\frac{\hat{\bar{V}}}{\bar{V}} - 1 + 1 \right)} \\ &= \frac{1}{\bar{V}} \times \frac{1}{\frac{\hat{\bar{V}}}{\bar{V}} - 1} \\ &= \frac{1}{\bar{V}} \left(1 - \frac{\hat{\bar{V}} - \bar{V}}{\bar{V}} + o\left(\frac{\hat{\bar{V}} - \bar{V}}{\bar{V}}\right) \right). \end{aligned}$$

Accordingly, this gives

$$\text{Var}\left[\frac{1}{\hat{\bar{V}}}\right] \approx \left(\frac{1}{\bar{V}}\right)^2 \times \frac{\text{Var}[\hat{\bar{V}}]}{\bar{V}^2}.$$

Lastly, we have

$$\text{Var}[\hat{\bar{V}}] \approx \bar{V}^4 \times \text{Var}\left[\frac{1}{\hat{\bar{V}}}\right],$$

or, with (8.6)

$$\text{Var}[\hat{\bar{V}}] \approx \bar{V}^4 \times \left(\frac{1}{m} - \frac{1}{T_p}\right) S_{1/u}^2. \quad (8.7)$$

By definition, variance $S_{1/u}^2$ is equal to

$$S_{1/u}^2 = \frac{1}{T_p - 1} \sum_{k \in U_p} \left(\frac{1}{u_k} - \frac{1}{\bar{V}} \right)^2. \quad (8.8)$$

Since quantity T_p is unknown, this relation can be estimated by

$$\frac{1}{m - 1} \sum_{k \in s_p} \left(\frac{1}{u_k} - \frac{1}{\bar{V}} \right)^2. \quad (8.9)$$

Given (8.7) and (8.9), we can easily determine the variance of estimator $\hat{\bar{V}}$ and consequently, that of estimator \hat{T}_p and lastly, that of the variable of interest \hat{Y} .

Comments 8.1. Estimator \hat{T}_p is biased and asymptotically unbiased.

Comment 8.2. If variables \hat{T}_p and \bar{y} are not independent then we would have

$$\begin{aligned} \text{Var}\left[\hat{T}_p \bar{y}\right] &= \bar{Y}^2 \text{Var}\left[\hat{T}_p\right] + T_p^2 \text{Var}[\bar{y}] \\ &\quad + \text{Var}\left[\hat{T}_p \bar{y}\right] \text{Var}[\bar{y}] \\ &\quad + \text{terms not linked to the} \\ &\quad \text{eventual non-independence} \\ &\quad \text{of the variables } \hat{T}_p \text{ and } \bar{y}. \end{aligned}$$

9. Numeric Illustration

A mechanical counter at a site in open country gives $T_v = 100$ vehicles. We assume that 20% of the vehicles have one person, 20% have two people, 20% have three people, 20% have four people and 20% have five people. This means there are 300 visitors to the site. The variance S_v^2 is equal to two disregarding finite population corrections. The average number of passengers \bar{V} is three. In effect, we have:

$$\begin{aligned} \frac{1}{\bar{V}} &= \frac{1}{1} \times \frac{20}{300} + \frac{1}{2} \times \frac{40}{300} + \frac{1}{3} \times \frac{60}{300} \\ &\quad + \frac{1}{4} \times \frac{80}{300} + \frac{1}{5} \times \frac{100}{300} = \frac{1}{3}. \end{aligned}$$

which gives $\bar{V} = 3$.

Let us now calculate an estimate of $S_{1/u}^2$. After simplifications of (8.8) and assuming that T_p is large enough compared to one, we have

$$S_{1/u}^2 \approx \frac{1}{T_p} \sum_{k \in U_p} \frac{1}{u_k^2} - \left(\frac{1}{\bar{V}}\right)^2.$$

Thus, we get

$$\begin{aligned} S_{1/u}^2 &= \frac{1}{30} \left(2 + 1 + \frac{2}{3} + \frac{1}{2} + \frac{2}{5} \right) - \frac{1}{3^2} \\ &= \frac{1}{30} \left(\frac{60 + 30 + 20 + 15 + 12}{30} \right) - \frac{1}{3^2} \\ &= \frac{137}{30^2} - \frac{1}{3^2} = \frac{37}{30^2}. \end{aligned}$$

Since we know $S_{1/u}^2$, we can calculate the variance of estimator \bar{V} . This gives

$$\text{Var}[\hat{\bar{V}}] \approx 3^4 \times \frac{37}{30^2} \times \frac{1}{m}.$$

Lastly, we can calculate the variance of estimator \hat{T}_p

$$\begin{aligned}\text{Var}[\hat{T}_p] &= T_p^2 \text{Var}[\hat{V}] \\ &= 10^4 \times 3^4 \times \frac{37}{30^2} \times \frac{1}{m}.\end{aligned}$$

The first approach gives a variance of estimator \hat{T}_p equal to

$$\text{Var}[\hat{T}_p] = 10^4 \times 2 \times \frac{1}{n}.$$

Thus, for estimator \hat{T}_p to have the same variance as estimator \hat{T}_p , size m of sample s_p must be equal to

$$m = 1.66n.$$

Our initial conclusion is that the second approach makes field operations simpler and less costly in terms of personnel because it only requires one enumerator. It is more accurate than a count that does not involve direct contact to obtain the composition of the tourist household. It requires only one sample about one and a half times larger than the first approach to produce the same accuracy, which is tolerable given the resulting simplification of collection. In practice, at all sites, the second approach will be the preferred application.

Conclusion

This article presented a broad description of a new method applicable to tourism statistics. It involves capturing tourists based on the consumption of certain services on which probabilistic samples are constructed. The weight share method makes it possible to shift from statistical accuracy of the services to the accuracy of the relevant tourism statistical units: the tour, the trip, the tourist household, the tourist or the tourist-night. However, the method requires numerous adaptations and complements to the weight share. We described one of these in detail, which is the estimate of the number of visitors to a site in open country. Two methods were tested. One, which was more accurate in terms of sample size, requires a relatively extensive organization and runs the risk of unacceptable errors in measurement. At the price of collecting slightly more data, the second method is preferred.

Other studies of this nature were conducted before and during the time of the survey so that it is difficult to present the full methodology in a single article.

Acknowledgements

The authors sincerely thank the two reviewers and associate editor who all made a significant contribution to improving the readability of this paper.

References

- Deville, J.-C. (1999). Les enquêtes par panel : En quoi diffèrent-elles des autres enquêtes ? suivi de : Comment attraper une population en se servant d'une autre. *Actes des journées de méthodologie statistiques, INSEE Méthodes*, 84-85-86, 63-82.
- Deville, J.-C., and Lavallée, P. (2006). Indirect Sampling: The Foundations of the Generalized Weight Share Method. *Survey Methodology*, 32, 2, 165-176.
- Deville, J.-C., Lavallée, P. and Maumy, M. (2005). Composition, factorisation et conditions d'optimalité (faible, forte) dans la méthode de partage des poids. Application à l'enquête sur le tourisme en Bretagne. *Actes des journées de méthodologie statistiques, INSEE Méthodes*.
- Hartley, H.O. (1962). Multiple Frame Surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.
- Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, Series C, 36, 99-118.
- Huygens, C. (1673). *Horologium Oscillatorium sive de motu pendulorum*.
- Lavallée, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method. *Survey Methodology*, 21, 25-32.
- Lavallée, P. (2002). Le sondage indirect, ou la méthode généralisée du partage des poids. Éditions de l'Université de Bruxelles, éditions Ellipses, Bruxelles.
- Lavallée, P., and Caron, P. (2001). Estimation using the generalised weight share method: The case of record linkage. *Survey Methodology*, 27, 155-169.
- Lund, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 282-288.
- Torres Manzanera, E., Sustacha Melijosa, I., Menéndez Estébanez, J.M. and Valdés Pelaáez, L. (2002). A solution to problems and disadvantages in statistical operations of surveys of visitors at accommodation establishments and at popular visitors places. (Éd. Ákos Probáld). *Proceedings Of The Sixth International Forum On Tourism Statistics. Hungarian Central Statistical Office, Budapest*.
- Valdés, L., De La Ballina, J., Aza, R., Loredó, E., Torres, E., Estébanez, J.M., Domínguez, J.S. and Del Valle, E. (2001). A methodology to measure tourism expenditure and total tourism production at the regional level. In *Tourism Statistics: International Perspectives and Current Issues*, (Ed. Lennon, J.J.), Continuum, Grande Bretagne, 317-334.

Combining Link-Tracing Sampling and Cluster Sampling to Estimate the Size of Hidden Populations: A Bayesian-Assisted Approach

Martín H. Félix-Medina and Pedro E. Monjardin¹

Abstract

Félix-Medina and Thompson (2004) proposed a variant of Link-tracing sampling in which it is assumed that a portion of the population, not necessarily the major portion, is covered by a frame of disjoint sites where members of the population can be found with high probabilities. A sample of sites is selected and the people in each of the selected sites are asked to nominate other members of the population. They proposed maximum likelihood estimators of the population sizes which perform acceptably provided that for each site the probability that a member is nominated by that site, called the nomination probability, is not small. In this research we consider Félix-Medina and Thompson's variant and propose three sets of estimators of the population sizes derived under the Bayesian approach. Two of the sets of estimators were obtained using improper prior distributions of the population sizes, and the other using Poisson prior distributions. However, we use the Bayesian approach only to assist us in the construction of estimators, while inferences about the population sizes are made under the frequentist approach. We propose two types of partly design-based variance estimators and confidence intervals. One of them is obtained using a bootstrap and the other using the delta method along with the assumption of asymptotic normality. The results of a simulation study indicate that (i) when the nomination probabilities are not small each of the proposed sets of estimators performs well and very similarly to maximum likelihood estimators; (ii) when the nomination probabilities are small the set of estimators derived using Poisson prior distributions still performs acceptably and does not have the problems of bias that maximum likelihood estimators have, and (iii) the previous results do not depend on the size of the fraction of the population covered by the frame.

Key Words: Bayesian approach; Capture-recapture; Design-based approach; Finite population; Hard-to-access population; Maximum likelihood; Model-based approach; Sampling frame.

1. Introduction

Link-tracing sampling (LTS) has been found appropriate for sampling hidden and hard-to-access human populations, such as drug-user, homeless-person, or illegal-worker populations. In this sampling method, an initial sample of people from the target population is selected, and the people in the initial sample are asked to nominate other members of the population. The nominated people who are not in the initial sample are included in the sample and they might be asked to nominate other persons. This process might continue until a specified stopping rule is satisfied (for a review of LTS, see Spreen 1992, and Thompson and Frank 2000).

Although LTS allows the sampler to make valid model-based inferences about a number of population parameters, in practical applications the assumptions about the initial sample are difficult to satisfy. (See Snijders 1992, Frank and Snijders 1994, and Heckathorn 2002). For instance, Frank and Snijders (1994) developed a variant of LTS in which the initial sample is a Bernoulli sample, that is, elements in the initial sample are included independently and with equal probabilities; however, in real studies the initial recruitment is generally carried out by using records of people obtained from health centers or police stations, and this induces a selection bias known as institutional bias.

The difficulty in satisfying, in practical situations, the assumptions about the initial sample motivated Félix-Medina and Thompson (2004) to develop a variant of LTS which does not require an initial Bernoulli sample. They assume that a portion, not necessarily the major portion, of the target population is covered by a sampling frame of accessible sites where members of the population can be found with high probability (for instance bars, hospitals, blocks or parks). A simple random sample of sites is selected, and the members that belong to each site are identified. Finally, as in ordinary LTS, the people in each site are asked to nominate other members of the population.

Those authors derived maximum likelihood estimators (MLEs) of the population sizes from probability models that describe both the number of elements found in each site and the probability that a member is nominated from a site, which is called the nomination probability. They also proposed model-based and partly design-based variance estimators, that is, estimators based on both the design used to select the initial sample and the assumed models. Throughout this paper we will call this type of estimator a "design-based-like" estimator. By a simulation study, the authors showed that the MLEs of the population sizes and their design-based-like variance estimators are robust to deviations from the assumed model, but that the model-based variance estimators are not robust. In addition, they

1. Martín H. Félix-Medina and Pedro E. Monjardin, Escuela de Ciencias Físico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacán Sinaloa, México.

found that the MLEs tend to seriously overestimate the population size if the nomination probabilities are small.

As indicated by those authors, the problem of overestimation that appears when the nomination probabilities are small is caused by the small amount of information contained in the sample, which is not enough to obtain stable estimates of the nomination probabilities. They suggest that a possible solution to this problem is to use the Bayesian approach to construct estimators that incorporate additional information about the population parameters.

In this work we use the Bayesian approach to assist us in the construction of estimators of the population sizes, while we make inferences under a frequentist approach. Thus, in addition to deriving point estimators we construct confidence intervals. For this purpose we use the strategy proposed by Félix-Medina and Thompson (2004) to construct confidence intervals based on the normal distribution and using design-based-like variance estimators obtained by the delta method. In addition, we construct design-based-like bootstrap confidence intervals. We have named this inferential approach “Bayesian-assisted”.

2. Sampling Design and Notation

The structure of the population and sampling design considered in this paper are the same as those proposed by Félix-Medina and Thompson (2004). A brief description of them follows. Let $U = \{u_1, \dots, u_\tau\}$ be a hidden human population of unknown size τ . Let U_1 be a subset of U formed by an unknown number τ_1 of people that can be found in different accessible sites, such as bars, parks, or blocks. Two assumptions about this sampling design are that a sampling frame of N of those sites can be constructed, and that the researcher has an operational rule which allows him or her to determine whether or not a person belongs to a site in the frame and, in the affirmative case, to locate that site. Notice that the subset U_1 covered by the frame is not assumed to be the major part of U and that, as in ordinary cluster sampling, a person in the frame is assumed to belong to only one site. Let A_i be the i -th site or cluster in the frame and m_i be the number of people who belong to A_i , $i = 1, \dots, N$; then $\tau_1 = \sum_{i=1}^N m_i$. Finally, let $U_2 = U - U_1$ be the portion of U not covered by the frame and let $\tau_2 = \tau - \tau_1$ be its size.

The sampling design is as follows. A sample $S_0 = \{A_1, \dots, A_n\}$ of n clusters is selected from the frame by simple random sampling without replacement, and the m_i persons who belong to each $A_i \in S_0$ are identified. Note that we have used the subscripts $1, \dots, n$ to denote the clusters in S_0 ; however, this does not mean that the first n clusters in the frame are necessarily the clusters in the sample. Next, the people in the sampled cluster A_i are asked to nominate

members in U , but only nominees in $U - A_i$ are considered. This procedure is repeated for every cluster $A_i \in S_0$. As a convention, we will say that a person is nominated by a cluster if he or she is nominated by at least one member of that cluster. Nominations from different clusters are carried out independently, and different nomination strategies can be used in different sites. For instance, in site A_i the m_i members, as a group, could carry out the nominations; whereas in another site A_j each of the m_j members could make nominations separately. Finally, for each nominee the researcher has to register the site or sites that nominated him or her, and the section U_1 or U_2 , to which the nominee belongs. Notice that this last piece of information could be obtained from the person who made the nomination or, if that is not possible, from an interview with the nominee.

The nomination of people by clusters will be indicated by the matrices $\mathbf{X}_1 = [x_{ij}^{(1)}]_{n \times \tau_1}$ and $\mathbf{X}_2 = [x_{ij}^{(2)}]_{n \times \tau_2}$, where $x_{ij}^{(1)} = 1$ if person $u_j \in U_1 - A_i$ is nominated by cluster A_i , and $x_{ij}^{(1)} = 0$ if $u_j \in A_i$ or u_j is not nominated by A_i . Similarly, $x_{ij}^{(2)} = 1$ if person $u_j \in U_2$ is nominated by cluster A_i , and $x_{ij}^{(2)} = 0$ otherwise. As noted by Félix-Medina and Thompson (2004), \mathbf{X}_1 and \mathbf{X}_2 are only known up to permutations of their columns because the people are not labelled. Therefore, inferences about τ_1 and τ_2 are based on the set of counts $\mathbf{y} = \{y_\omega\}$, where y_ω , $\omega \subseteq \Omega = \{1, \dots, n\}$, $\omega \neq \emptyset$, indicates the number of people in U who are nominated by every sampled cluster A_i with i in the set ω , but not otherwise. For instance, if $\omega = \{4, 7, 8\}$, y_ω would be the number of people in U who are nominated by only A_4 , A_7 and A_8 .

3. Estimators of the Population Sizes Based on Posterior Modes

Félix-Medina and Thompson noted the resemblance between their sampling design and that of multiple capture-recapture sampling (MCRS). This makes it possible to apply to our case some of the Bayesian models that have been proposed for analyzing MCRS. See Fienberg, Johnson and Junker (1999) for a review of Bayesian analyses of MCRS. In this work, we use a model considered by Castledine (1981) for the prior distributions of the logits of the nomination probabilities, along with some models for the prior distributions of the population sizes.

As in Félix-Medina and Thompson (2004), we will suppose that the sizes m_1, \dots, m_N of the clusters A_1, \dots, A_N are realizations of independent Poisson random variables M_1, \dots, M_N with mean λ_1 . We will denote by $p_i^{(k)}$ the probability that a person in $U_i - A_i$ is nominated by the site $A_i \in S_0$. The probabilities $p_i^{(k)}$ will be called nomination probabilities. In addition, we will suppose that conditionally on the sizes m_1, \dots, m_n of the clusters in S_0 , on τ_1 and τ_2 ,

and on the $p_i^{(k)}$'s, the variables $x_{ij}^{(k)}$ are realizations of independent Bernoulli random variables $X_{ij}^{(k)}$ with means $p_i^{(k)}$, $i = 1, \dots, n$ and $k = 1, 2$.

Félix-Medina and Thompson (2004) used the fact that the joint conditional distribution of $(M_1, \dots, M_n, \tau_1 - \sum_1^n M_i)$, given that $\sum_1^n m_i = \tau_1$, is a multinomial distribution with parameters τ_1 and $(1/N, \dots, 1/N, 1 - n/N)$, and applied a procedure used by Darroch (1958) to show that the likelihood function of $\tau_1, \tau_2, \mathbf{p}_1 = \{p_i^{(1)}\}_1^n$ and $\mathbf{p}_2 = \{p_i^{(2)}\}_1^n$ is the product of the following factors:

$$\begin{aligned} f(\mathbf{m}_s | \tau_1) &= \frac{\tau_1!}{(\tau_1 - m)! \prod_1^n m_i!} (1/N)^m (1 - n/N)^{\tau_1 - m} \\ f(\mathbf{y}^{(1-0)} | \mathbf{m}_s, \tau_1, \mathbf{p}_1) &= \frac{(\tau_1 - m)!}{(\tau_1 - m - r_1)! \prod_{\omega \neq \emptyset} y_\omega^{(1-0)}!} \prod_{i=1}^n [P_i^{(1)}]^{z_i^{(1-0)}} \\ &\quad \times [1 - p_i^{(1)}]^{\tau_1 - m - z_i^{(1-0)}} \\ f(\mathbf{y}^{(A_1)}, \dots, \mathbf{y}^{(A_n)} | \mathbf{m}_s, \mathbf{p}_1) &= \prod_{i=1}^n \frac{m_i!}{(m_i - w_i)! \prod_{\omega \neq \emptyset} y_\omega^{(A_i)}!} [P_i^{(1)}]^{z_i^{(A_i)}} \\ &\quad \times [1 - p_i^{(1)}]^{m - z_i^{(A_i)}} \\ f(\mathbf{y}^{(2)} | \mathbf{m}_s, \tau_2, \mathbf{p}_2) \\ &= \frac{\tau_2!}{(\tau_2 - r_2)! \prod_{\omega \neq \emptyset} y_\omega^{(2)}!} \prod_{i=1}^n [P_i^{(2)}]^{z_i^{(2)}} [1 - p_i^{(2)}]^{\tau_2 - z_i^{(2)}}, \end{aligned}$$

where $\mathbf{m}_s = \{m_i\}_1^n$, $m = \sum_1^n m_i$ is the observed value of the random variable M that indicates the number of people in S_0 ; $\mathbf{y}^{(1-0)} = \{y_\omega^{(1-0)}\}_{\omega \neq \emptyset}$, $\mathbf{y}^{(2)} = \{y_\omega^{(2)}\}_{\omega \neq \emptyset}$, and $\mathbf{y}^{(A_i)} = \{y_\omega^{(A_i)}\}_{\omega \neq \emptyset}$, $A_i \in S_0$, are the sets of counts obtained from \mathbf{y} , that correspond to the counts of nominated people in $U_1 - S_0$, U_2 and $A_i \in S_0$, respectively; $z_i^{(0)} = \sum_{j \neq i} \sum_{\omega \supset i} y_\omega^{(A_j)}$, $z_i^{(1-0)} = \sum_{\omega \supset i} y_\omega^{(1-0)}$ and $z_i^{(2)} = \sum_{\omega \supset i} y_\omega^{(2)}$ are the observed values of the random variables $Z_i^{(0)}$, $Z_i^{(1-0)}$ and $Z_i^{(2)}$ that indicate the numbers of distinct people in S_0 , $U_1 - S_0$ and U_2 , respectively, that are nominated by A_i ; and $r_1 = \sum_{\omega \neq \emptyset} y_\omega^{(1-0)}$, $r_2 = \sum_{\omega \neq \emptyset} y_\omega^{(2)}$ and $w_i = \sum_{\omega \neq \emptyset} y_\omega^{(A_i)}$ are the observed values of the random variables R_1 , R_2 and W_i that indicate the numbers of distinct people in $U_1 - S_0$, U_2 and A_i , respectively, that are nominated by at least one of the clusters in S_0 .

We will now focus on the problem of defining the prior distributions of $\tau_1, \tau_2, \mathbf{p}_1$ and \mathbf{p}_2 . In the case of τ_1 and τ_2 , we will consider the following three models for the prior distributions:

Poisson-Gamma Distributions

$$\begin{aligned} \pi(\tau_1 | \lambda_1) &\propto (N\lambda_1)^{\tau_1} / \tau_1! \text{ and } \pi(\lambda_1) \propto \lambda_1^{a_1 - 1} e^{-b_1 \lambda_1}, \\ \pi(\tau_2 | \lambda_2) &\propto \lambda_2^{\tau_2} / \tau_2! \text{ and } \pi(\lambda_2) \propto \lambda_2^{a_2 - 1} e^{-b_2 \lambda_2}, \end{aligned}$$

where a_1, b_1, a_2, b_2 are known constants, and (τ_1, λ_1) and (τ_2, λ_2) are independent.

Jeffreys' Distributions

$\pi(\tau_k) \propto 1/\tau_k$, where $k = 1, 2$, and τ_1 and τ_2 are independent random variables.

Uniform Distributions

$\pi(\tau_k) \propto 1$, where $k = 1, 2$, and τ_1 and τ_2 are independent random variables.

The prior Poisson distribution of τ_1 defined in the first case is motivated by the fact that $\tau_1 = \sum_1^n M_i$, and that M_i is a Poisson variable with mean λ_1 . Notice that this case allows the researcher to use information about τ_1 and τ_2 which is known prior to the observation of the sample. On the other hand, the distributions defined in the other two cases are not informative.

In the case of the nomination probabilities $p_i^{(k)}$'s, following Castledine (1981), we will suppose that the $p_i^{(k)}$'s are exchangeable and will use his two-stage normal model for the logits $\alpha_i^{(k)} = \log[p_i^{(k)} / (1 - p_i^{(k)})]$ of the $p_i^{(k)}$'s:

$$\alpha_i^{(k)} | \theta_k \sim N(\theta_k, \sigma_k^2),$$

$$\text{and } \theta_k \sim N(\mu_k, \gamma_k^2); i = 1, \dots, n, k = 1, 2, \quad (1)$$

where $N(\theta_k, \sigma_k^2)$ stands for the normal distribution with mean θ_k and variance σ_k^2 ; σ_k^2, μ_k and γ_k^2 are known constants; and the $\alpha_i^{(k)}$'s are conditionally independent given θ_k . Under the assumption of exchangeability the $\alpha_i^{(k)}$'s are not independent, but information about any one of them is used to obtain information about any other of the $\alpha_i^{(k)}$'s. Of course, if we wanted independent priors for the $\alpha_i^{(k)}$'s, we could obtain a one-stage normal model from (1) by setting $\theta_k = \mu_k$ and $\gamma_k^2 = 0$, $k = 1, 2$.

Finally, we will suppose that all the random vectors (τ_k, λ_k) and (α_k, θ_k) , where $\alpha_k = (\alpha_1^{(k)}, \dots, \alpha_n^{(k)})$, $k = 1, 2$, are mutually independent.

Although we defined three types of prior distributions for τ_1 and τ_2 , they can be treated in a unified way because the prior marginal distributions of τ_1 and τ_2 , obtained from the Poisson-Gamma distributions, are the Negative binomial distributions:

$$\pi(\tau_1) \propto \frac{\Gamma(\tau_1 + a_1)}{\tau_1!} \left(\frac{N}{N + b_1} \right)^{\tau_1} \quad (2)$$

$$\text{and } \pi(\tau_2) \propto \frac{\Gamma(\tau_2 + a_2)}{\tau_2!} \left(\frac{1}{1 + b_2} \right)^{\tau_2},$$

where $\Gamma(\cdot)$ denotes the Gamma function. The Jeffreys' and Uniform distributions are limiting cases of (2) obtained by making $a_k = b_k = 0$, $k = 1, 2$, and $a_k = 1$, $b_k = 0$, $k = 1, 2$,

respectively. Note that the Gamma distribution is not defined for these values of a_k and b_k ; however, for the derivation of the estimators we can use these values in (2).

The posterior joint distribution of τ_1, τ_2, α_1 , and α_2 can be expressed as

$$\begin{aligned} & \pi(\tau_1, \tau_2, \alpha_1, \alpha_2 | \text{data}) \\ & \propto \frac{(N-n)^{\tau_1} \Gamma(\tau_1 + a_1)}{(\tau_1 - m - r_1)!(N + b_1)^{\tau_1}} \prod_{i=1}^n \frac{\exp[\alpha_i^{(1)} z_i^{(1)}]}{[1 + \exp[\alpha_i^{(1)}]]^{\tau_1 - m_i}} \\ & \times \exp \left[-\frac{\sum_{i=1}^n (\alpha_i^{(1)} - \bar{\alpha}^{(1)})^2}{2\sigma_1^2} - \frac{\Gamma(\tau_2 + a_2)}{(\tau_2 - r_2)!(b_2 + 1)^{\tau_2}} - \frac{(\bar{\alpha}^{(1)} - \mu_1)^2}{2\nu_1} \right] \\ & \times \prod_{i=1}^n \frac{\exp[\alpha_i^{(2)} z_i^{(2)}]}{[1 + \exp[\alpha_i^{(2)}]]^{\tau_2}} \exp \left[-\frac{\sum_{i=1}^n (\alpha_i^{(2)} - \bar{\alpha}^{(2)})^2}{2\sigma_2^2} - \frac{(\bar{\alpha}^{(2)} - \mu_2)^2}{2\nu_2} \right] \end{aligned} \quad (3)$$

where $z_i^{(1)} = z_i^{(0)} + z_i^{(1-0)}$ is the observed value of the random variable $Z_i^{(1)} = Z_i^{(0)} + Z_i^{(1-0)}$ that indicates the number of distinct people in U_1 , either in S_0 or in $U_1 - S_0$, that are nominated by A_i ; $\bar{\alpha}^{(k)}$ is the arithmetic mean of the $\alpha_i^{(k)}$; and $\nu_k = \nu_k^2 + \sigma_k^2/n$, $k = 1, 2$.

Since we cannot compute the analytical integral of (3) with respect to $\alpha_i^{(1)}$ and $\alpha_i^{(2)}$, we will not try to obtain expressions for the posterior distributions of τ_1 and τ_2 , but, as in Castledine (1981), we will use the mode of $\pi(\tau_1, \tau_2, \alpha_1, \alpha_2 | \text{data})$ as an estimator of $(\tau_1, \tau_2, \alpha_1, \alpha_2)$. Using this strategy, we have that the proposed estimator is the solution to the system of equations:

$$\begin{aligned} \hat{\tau}_1 &= \frac{M + R_1 + (1 - n/N)[N(a_1 - 1)/(N + b_1)] \prod_{i=1}^n (1 - \hat{p}_i^{(1)})}{1 - (1 - n/N)[N/(N + b_1)] \prod_{i=1}^n (1 - \hat{p}_i^{(1)})}; \\ \hat{p}_i^{(1)} &= \frac{\exp\{\hat{\alpha}_i^{(1)}\}}{1 + \exp\{\hat{\alpha}_i^{(1)}\}} = \frac{Z_i^{(1)}}{\hat{\tau}_1 - M_i} - \frac{\hat{\alpha}_i^{(1)} - \hat{\bar{\alpha}}^{(1)}}{(\hat{\tau}_1 - M_i)\sigma_1^2} \\ & - \frac{\hat{\bar{\alpha}}^{(1)} - \mu_1}{n(\hat{\tau}_1 - M_i)\nu_1}; i = 1, \dots, n; \end{aligned} \quad (4)$$

$$\begin{aligned} \hat{\tau}_2 &= \frac{R_2 + [(a_2 - 1)/(1 + b_2)] \prod_{i=1}^n (1 - \hat{p}_i^{(2)})}{1 - [1/(1 + b_2)] \prod_{i=1}^n (1 - \hat{p}_i^{(2)})}; \\ \hat{p}_i^{(2)} &= \frac{\exp\{\hat{\alpha}_i^{(2)}\}}{1 + \exp\{\hat{\alpha}_i^{(2)}\}} = \frac{Z_i^{(2)}}{\hat{\tau}_2} - \frac{\hat{\alpha}_i^{(2)} - \hat{\bar{\alpha}}^{(2)}}{\hat{\tau}_2 \sigma_2^2} \\ & - \frac{\hat{\bar{\alpha}}^{(2)} - \mu_2}{n\hat{\tau}_2 \nu_2}; i = 1, \dots, n; \end{aligned} \quad (5)$$

where $\hat{\bar{\alpha}}^{(k)} = \sum_{i=1}^n \hat{\alpha}_i^{(k)} / n$, $k = 1, 2$. From this, an estimator of τ is $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$.

The forms of these estimators are basically adjustments to the forms of the MLE's proposed by Félix-Medina and Thompson (2004) so that the proposed estimators incorporate the initial information about τ_k and $\alpha_i^{(k)}$, $i = 1, \dots, n$; $k = 1, 2$. Also, as a referee has noted, the estimator $\hat{p}_i^{(k)}$ has the form of the MLE of $p_i^{(k)}$ followed by shrinkage terms, one of $\alpha_i^{(k)}$ toward the arithmetic mean $\hat{\bar{\alpha}}^{(k)}$, and another of $\hat{\bar{\alpha}}^{(k)}$ toward the prior mean μ_k .

4. Confidence Intervals for the Population Sizes

As was indicated earlier, we will use the frequentist approach to obtain design-based-like confidence intervals that are robust to deviations from the assumed Poisson distribution of the M_i 's. We will consider bootstrap intervals and Wald intervals based on a normal approximation (see Agresti 2002, page 13 and Evans, Kim and O'Brien 1996 for the latter terminology).

4.1 Bootstrap Confidence Intervals

We will use a version of the bootstrap obtained by combining the bootstrap variant for finite populations proposed by Booth, Butler and Hall (1994) and the parametric bootstrap variant (see Davison and Hinkley 1997, Chapter 2).

The steps of the procedure that we propose are the following. (i) Construct an artificial population of N values of m_i 's by repeating N/n times, assuming that N/n is an integer, the selected sample of n cluster sizes m_1, \dots, m_n . If $N = kn + r$, where k and r are positive integers, construct the population by repeating k times the selected sample of n cluster sizes and add to this set of m_i 's a simple random sample without replacement (SRSWOR) of r values of m_i 's selected from the observed sample of n cluster sizes. (ii) Select a SRSWOR of size n from the population of the m_i 's. Let i_1, \dots, i_n be the indices of the m_i 's in the sample. (iii) For each $i = i_1, \dots, i_n$, draw samples of sizes $\hat{\tau}_1 - m_i$ and $\hat{\tau}_2$ from Bernoulli distributions with means $\hat{p}_i^{(1)}$ and $\hat{p}_i^{(2)}$, respectively, where $\hat{\tau}_1, \hat{\tau}_2, \hat{p}_i^{(1)}$ and $\hat{p}_i^{(2)}$ are

the estimates of $\tau_1, \tau_2, p_i^{(1)}$ and $p_i^{(2)}$ computed from the original observed sample. These samples simulate the values of the sets $\{x_{ij}^{(1)}\}$ and $\{x_{ij}^{(2)}\}$ of indicator variables. (iv) Compute estimates of τ_1, τ_2 and τ from the samples drawn in steps (ii) and (iii) using the same procedure as that used to compute the original estimates $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$. (v) Obtain the bootstrap distributions of $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$ by repeating (i)–(iv) a large number B of times, and computing the empirical distributions from the sets of B values of $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$. (vi) Construct the $100(1-\alpha)\%$ bootstrap confidence intervals for τ_1, τ_2 and τ by using either the basic or the percentile method (see Davison and Hinkley 1997, Chapter 5, for descriptions of these methods). In the basic method the interval for τ is $[2\hat{\tau} - \hat{\tau}^{(1-\alpha/2)}, 2\hat{\tau} - \hat{\tau}^{(\alpha/2)}]$, and in the percentile method it is $[\hat{\tau}^{(\alpha/2)}, \hat{\tau}^{(1-\alpha/2)}]$, where $\hat{\tau}^{(\alpha/2)}$ and $\hat{\tau}^{(1-\alpha/2)}$ are the lower and upper $\alpha/2$ points of the bootstrap distribution of the original estimate $\hat{\tau}$ of τ .

Note that this variant of the bootstrap does not use the assumed Poisson distribution of the M_i 's, but it uses the sampling design employed to select the initial sample of clusters. Thus, we can consider that the resulting confidence intervals are robust to deviations from the assumed distribution of the M_i 's.

If bootstrap estimates of the variances of $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$ were also desired, simple estimates could be obtained by computing the sample variances of the sets of B values of those estimators.

4.2 Wald Confidence Intervals

Though in this work we will not justify theoretically that the proposed estimators of the population sizes are asymptotically normally distributed, we will suppose that the normal distribution is a reasonable approximation to the distributions of the estimators. Thus, we will construct $100(1-\alpha)\%$ design-based-like Wald confidence intervals for the population sizes, which have the form $\hat{\tau}_k \pm z_{1-\alpha/2} \sqrt{\hat{V}(\hat{\tau}_k)}$, where $z_{1-\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution, and $\hat{V}(\hat{\tau}_k)$ is a design-based-like estimator of the variance of $\hat{\tau}_k$.

To construct this type of interval, we will firstly derive design-based-like variance estimators by applying the same strategy as that used by Félix-Medina and Thompson (2004). In that strategy, the distribution of the cluster sizes is not employed, but it is replaced by the distribution of the sampling design used to select the initial sample S_0 . This is carried out by means of the formula:

$$\mathbf{V}(\hat{\tau}_k) = \mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_k | \mathbf{m}_s)] + \mathbf{E}_p[\mathbf{V}_\xi(\hat{\tau}_k | \mathbf{m}_s)], \quad (6)$$

where $\mathbf{E}_\xi(\hat{\tau}_k | \mathbf{m}_s)$ and $\mathbf{V}_\xi(\hat{\tau}_k | \mathbf{m}_s)$ denote the conditional model-based expectation and variance operators, given that $\mathbf{M}_s = \mathbf{m}_s$; and $\mathbf{E}_p(\cdot)$ and $\mathbf{V}_p(\cdot)$ denote the design-based expectation and variance operators. Thus, the variance

estimators are obtained by applying (6) to the first-order Taylor's approximations $\hat{\tau}_1^*$ and $\hat{\tau}_2^*$ of $\hat{\tau}_1$ and $\hat{\tau}_2$, respectively, about the model-based expectations of $c_s^{(1)} = (\mathbf{M}_s, \mathbf{Z}_s^{(1)}, R_1)$ and $c_s^{(2)} = (\mathbf{Z}_s^{(2)}, R_2)$, where $\mathbf{Z}_s^{(k)} = (Z_1^{(k)}, \dots, Z_n^{(k)})$, $k = 1, 2$.

Using the previously described strategy, and the fact that $Z_i^{(1)} | \mathbf{m}_s \sim \text{bin}(\tau_1 - m_i, p_i^{(1)})$ and $R_1 | \mathbf{m}_s \sim \text{bin}(\tau_1 - m, 1 - Q_1)$, where $Q_1 = \prod_{i=1}^n (1 - p_i^{(1)})$, we have that an estimator of $\mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_1^* | \mathbf{m}_s)]$ is

$$\hat{V}_{11} = n(1 - n/N) \hat{K}^2 \frac{1}{n-1} \sum_{i=1}^n (m_i - \bar{m})^2, \quad (7)$$

where $\bar{m} = n^{-1} \sum_{i=1}^n m_i$; $\hat{K} = \hat{Q}_1 / [\hat{A}_1(\hat{\tau}_1 - m - r_1)]$; $\hat{Q}_1 = \prod_{i=1}^n (1 - \hat{p}_i^{(1)})$;

$$\hat{A}_1 = \sum_{i=1}^n \frac{(\hat{p}_i^{(1)})^2}{\hat{B}_i^{(1)}} - \hat{C}_1 + \frac{1}{\hat{\tau}_1 + a_1 - 1} - \frac{1}{\hat{\tau}_1 - m - r_1};$$

$$\hat{B}_i^{(1)} = (\hat{\tau}_1 - m_i) \hat{p}_i^{(1)} (1 - \hat{p}_i^{(1)}) + \sigma_1^{-2}, i = 1, \dots, n;$$

and

$$\hat{C}_1 = \frac{(v_1^{-1} - n\sigma_1^{-2}) \left[n^{-1} \sum_{i=1}^n \hat{p}_i^{(1)} / \hat{B}_i^{(1)} \right]^2}{1 + n^{-1} (v_1^{-1} - n\sigma_1^{-2}) n^{-1} \sum_{i=1}^n 1 / \hat{B}_i^{(1)}}. \quad (8)$$

In addition, since $\text{Cov}(Z_i^{(1)}, R_1 | \mathbf{m}_s) = (\tau_1 - m) Q_1 p_i^{(1)}$, an estimator of $\mathbf{E}_p[\mathbf{V}_\xi(\hat{\tau}_1^* | \mathbf{m}_s)]$ is

$$\hat{V}_{12} = \hat{A}_1^{-2} \left\{ \sum_{i=1}^n \left(\frac{\hat{p}_i^{(1)} - \hat{D}_1}{\hat{B}_i^{(1)}} \right)^2 (\hat{\tau}_1 - m_i) \hat{p}_i^{(1)} (1 - \hat{p}_i^{(1)}) + \frac{(\hat{\tau}_1 - m) \hat{Q}_1 (1 - \hat{Q}_1)}{(\hat{\tau}_1 - m - r_1)^2} - \frac{2(\hat{\tau}_1 - m) \hat{Q}_1}{\hat{\tau}_1 - m - r_1} \sum_{i=1}^n \left(\frac{\hat{p}_i^{(1)} - \hat{D}_1}{\hat{B}_i^{(1)}} \right) \hat{p}_i^{(1)} \right\}, \quad (9)$$

where

$$\hat{D}_1 = \frac{n^{-1} (v_1^{-1} - n\sigma_1^{-2}) n^{-1} \sum_{i=1}^n \hat{p}_i^{(1)} / \hat{B}_i^{(1)}}{1 + n^{-1} (v_1^{-1} - n\sigma_1^{-2}) n^{-1} \sum_{i=1}^n 1 / \hat{B}_i^{(1)}}.$$

Therefore, a design-based-like estimator of $\mathbf{V}(\hat{\tau}_1)$ is $\hat{V}(\hat{\tau}_1) = \hat{V}_{11} + \hat{V}_{12}$.

In the case of $\hat{\tau}_2^*$, since $Z_i^{(2)} | \mathbf{m}_s \sim \text{bin}(\tau_2, p_i^{(2)})$ and $R_2 | \mathbf{m}_s \sim \text{bin}(\tau_2, 1 - Q_2)$, where $Q_2 = \prod_{i=1}^n (1 - p_i^{(2)})$, it follows that $\mathbf{E}_\xi(\hat{\tau}_2^* | \mathbf{m}_s)$ does not depend on \mathbf{m}_s , and consequently that $\mathbf{V}_p[\mathbf{E}_\xi(\hat{\tau}_2^* | \mathbf{m}_s)] \approx 0$. Therefore, since $\text{Cov}(Z_i^{(2)}, R_2 | \mathbf{m}_s) = \tau_2 Q_2 p_i^{(2)}$, an estimator of $\mathbf{V}(\hat{\tau}_2)$ is

$$\hat{\mathbf{V}}(\hat{\tau}_2) = \hat{A}_2^{-2} \left\{ \begin{aligned} & \sum_{i=1}^n \left(\frac{\hat{p}_i^{(2)} - \hat{D}_2}{\hat{B}_i^{(2)}} \right)^2 \hat{\tau}_2 \hat{p}_i^{(2)} (1 - \hat{p}_i^{(2)}) \\ & + \frac{\hat{\tau}_2 \hat{Q}_2 (1 - \hat{Q}_2)}{(\hat{\tau}_2 - r_2)^2} \\ & - \frac{2 \hat{\tau}_2 \hat{Q}_2}{\hat{\tau}_2 - r_2} \sum_{i=1}^n \left(\frac{\hat{p}_i^{(2)} - \hat{D}_2}{\hat{B}_i^{(2)}} \right) \hat{p}_i^{(2)} \end{aligned} \right\} \quad (10)$$

where $\hat{Q}_2 = \prod_{i=1}^n (1 - \hat{p}_i^{(2)})$,

$$\hat{A}_2 = \sum_{i=1}^n \frac{(\hat{p}_i^{(2)})^2}{\hat{B}_i^{(2)}} - \hat{C}_2 + \frac{1}{\hat{\tau}_2 + a_2 - 1} - \frac{1}{\hat{\tau}_2 - r_2},$$

$$\hat{B}_i^{(2)} = \hat{\tau}_2 \hat{p}_i^{(2)} (1 - \hat{p}_i^{(2)}) + \sigma_2^{-2}, \quad i = 1, \dots, n,$$

$$\hat{C}_2 = \frac{(v_2^{-1} - n\sigma_2^{-2}) \left[n^{-1} \sum_{i=1}^n \hat{p}_i^{(2)} / \hat{B}_i^{(2)} \right]^2}{1 + n^{-1} (v_2^{-1} - n\sigma_2^{-2}) n^{-1} \sum_{i=1}^n 1 / \hat{B}_i^{(2)}},$$

and

$$\hat{D}_2 = \frac{n^{-1} (v_2^{-1} - n\sigma_2^{-2}) n^{-1} \sum_{i=1}^n \hat{p}_i^{(2)} / \hat{B}_i^{(2)}}{1 + n^{-1} (v_2^{-1} - n\sigma_2^{-2}) n^{-1} \sum_{i=1}^n 1 / \hat{B}_i^{(2)}}.$$

Finally, since the no dependency of $\mathbf{E}_\xi(\hat{\tau}_2^* | \mathbf{m}_s)$ on \mathbf{m}_s implies that $\mathbf{Cov}(\hat{\tau}_1^*, \hat{\tau}_2^*) \approx 0$, it follows that a variance estimator of $\hat{\tau}$ is $\hat{\mathbf{V}}(\hat{\tau}) = \hat{\mathbf{V}}(\hat{\tau}_1) + \hat{\mathbf{V}}(\hat{\tau}_2)$.

5. Monte Carlo Study

We considered four populations; a description of each one is presented in Table 1. In the pair formed by Populations I and II the frame covered about 45% of the population, whereas in the pair formed by Populations III and IV the frame covered about 70% of the population. The populations of each pair were very similar, except that in one of the populations of each pair the distribution of the M_i 's was Poisson, whereas in the other it was Negative Binomial. The nomination probabilities $p_i^{(k)}$, $i = 1, \dots, N$, $k = 1, 2$, were generated using the model $p_i^{(k)} = 1 - \exp(-\beta_k m_i)$, where the values of β_k were set so that the following values of $\bar{p}^{(k)} = \sum_{i=1}^N p_i^{(k)} / N$ were obtained. For Populations I and II: $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0.05, 0.01)$ and $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0.01, 0.002)$. For Populations III and IV: $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0.05, 0.03)$ and $(\bar{p}^{(1)}, \bar{p}^{(2)}) \approx (0.01, 0.006)$. The model employed to generate the $p_i^{(k)}$'s is a model used in catch-effort methods (see Seber 1982, Chapter 7 for a description of those methods). As an associate editor has noted, this model implies that the number of people nominated by cluster A_i has expectation $(\tau_1 - m_i)(1 - \exp(-\beta_1 m_i)) + \tau_2(1 - \exp(-\beta_2 m_i))$, and consequently

the number of nominated people is approximately proportional to m_i . Notice that the assumed exchangeable model for $p_i^{(k)}$ does not entail such a relationship with m_i . Since the estimation of $p_i^{(k)}$ depends mainly on $z_i^{(k)}$, the number of people in U_k nominated by the cluster A_i , we expect the omission of this relationship not to affect the efficiency of the estimator of $p_i^{(k)}$. Darroch (1958) has shown, in the case of maximum likelihood estimation, that no significant gain is obtained by assuming the catch-effort model.

Table 1
Parameters of Simulated Populations

Population I	Population II	Population III	Population IV
$N = 250$	$N = 250$	$N = 250$	$N = 250$
M_i Poisson	M_i Neg. Binomial	M_i Poisson	M_i Neg. Binomial
$\mathbf{E}(M_i) = 7.2$	$\mathbf{E}(M_i) = 7.2$	$\mathbf{E}(M_i) = 7.2$	$\mathbf{E}(M_i) = 7.2$
$\mathbf{V}(M_i) = 7.2$	$\mathbf{V}(M_i) = 24.48$	$\mathbf{V}(M_i) = 7.2$	$\mathbf{V}(M_i) = 24.48$
$\tau_1 = 1,811$	$\tau_1 = 1,872$	$\tau_1 = 1,811$	$\tau_1 = 1,872$
$\tau_2 = 2,200$	$\tau_2 = 2,200$	$\tau_2 = 700$	$\tau_2 = 700$
$\tau = 4,011$	$\tau = 4,072$	$\tau = 2,511$	$\tau = 2,572$
$\tau_1 / \tau = 0.45$	$\tau_1 / \tau = 0.46$	$\tau_1 / \tau = 0.72$	$\tau_1 / \tau = 0.73$

For Populations I and II the values of the parameters of the prior distributions were $\sigma_k^2 = 25$, $\mu_k = -3.5$, $\gamma_k^2 = 25$, $k = 1, 2$, $a_1 = 1$, $b_1 = 0.1$, $a_2 = 7.84$, $b_2 = 0.0028$, so that $\mathbf{E}(\lambda_1) = 10$, $\mathbf{V}(\lambda_1) = 100$, $\mathbf{E}(\lambda_2) = 2,800$, and $\mathbf{V}(\lambda_2) = 10^6$. For Populations III and IV the values of the parameters were $\sigma_k^2 = 9$, $\mu_k = -3.5$, $\gamma_k^2 = 9$, $k = 1, 2$, $a_1 = 1$, $b_1 = 0.1$, $a_2 = 8$, $b_2 = 0.01$, so that $\mathbf{E}(\lambda_1) = 10$, $\mathbf{V}(\lambda_1) = 100$, $\mathbf{E}(\lambda_2) = 800$, and $\mathbf{V}(\lambda_2) = 80,000$. These values imply that the prior distributions are well dispersed over relatively large intervals that contain the parameters of interest.

The simulation experiment was carried out as follows. From each population of $N = 250$ values of m_i 's, a SRSWOR of $n = 25$ values was selected. From cluster A_i in the sample, the values of $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ were generated by drawing samples of sizes $\tau_1 - m_i$ and τ_2 from Bernoulli distributions with means $p_i^{(1)}$ and $p_i^{(2)}$, respectively. These data were used to compute the following estimators of the population sizes: the set of MLEs $\hat{\tau}_1, \hat{\tau}_2$, and $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$ proposed by Félix-Medina and Thompson (2004); and the three sets of Bayesian estimators $\hat{\tau}_1^a, \hat{\tau}_2^a$, and $\hat{\tau}^a = \hat{\tau}_1^a + \hat{\tau}_2^a$, $a = U, J, P$, obtained by using as prior distributions the Uniform (U), Jeffreys' (J), and Poisson (P) distributions, respectively. In addition, variance estimators and confidence intervals were also computed. Bootstrap intervals were computed by the basic method, with the exception of the intervals based on the estimators $\hat{\tau}_1^P, \hat{\tau}_2^P$ and $\hat{\tau}^P$, which were computed by the percentile method. All bootstrap estimators were obtained by using 2,000 bootstrap samples. Finally, the performance of the point and

interval estimators was evaluated by using $r = 10,000$ trials of the previous procedure.

The performance of an estimator $\hat{\tau}$, say, was evaluated by its relative bias and the square root of its relative mean square error, defined as $r - \text{bias} = \sum_i (\hat{\tau}_i - \tau)/(r\tau)$ and $\sqrt{r - \text{mse}} = \sqrt{\sum_i (\hat{\tau}_i - \tau)^2/(r\tau^2)}$, where $\hat{\tau}_i$ was the value of $\hat{\tau}$ obtained in the i -th trial. The performance of a variance estimator was also evaluated by its relative bias and the square root of its relative mean square error, which were similarly defined to those of an estimator of the population size, but using the empirically determined variance instead of the real variance. Finally, the performance of the 95% confidence intervals was evaluated by their coverage probabilities and their average lengths.

6. Results and Discussion

Because of restrictions of space, in Tables 2 to 4 we present only a selection of the results of the numerical study. However, the next comments refer to the complete set of results.

Despite the limitations of the simulation study, we can conclude that the main factor that affects the performance of the estimators and confidence intervals is the magnitude of the $p_i^{(k)}$'s. When they are large and regardless of the distribution of the M_i 's and the size of the fraction τ_1/τ covered by the frame, every one of the estimators of the τ 's and design-based-like confidence intervals (Wald or bootstrap) perform satisfactorily. However, when the $p_i^{(k)}$'s

are small and in spite of all the other factors, only the Bayesian estimators $\hat{\tau}_k^P$ perform acceptably. It is worth noting that when the $p_i^{(k)}$'s are small, the Bayesian estimators $\hat{\tau}_k^U$ and $\hat{\tau}_k^J$ perform better than the MLE's $\hat{\tau}_k$; however, the performance of $\hat{\tau}_k^U$ and $\hat{\tau}_k^J$ is not good enough to make reliable inferences.

Bootstrap confidence intervals for τ_1 based on $\hat{\tau}_1^P$ did not perform as well as Wald intervals when the $p_i^{(k)}$'s were small or the M_i 's were not Poisson distributed. The explanation of this result and the development of better bootstrap intervals are topics that require further research.

Finally, the best performance of the set of estimators $\hat{\tau}_k^P$ is a consequence of the greater amount of information used by them. Though we used relatively flat prior distributions for the τ_k 's, the information supplied by them was enough to avoid the problems of bias and high variability observed in the other estimators. We carried out some additional simulation trials, and the results (which are not reported in the tables) indicate that, as long as the prior distributions are kept relatively flat, the estimates are not affected by the values of the parameters of the prior distributions. Obviously, misleading initial information combined with small values of the $p_i^{(k)}$'s will affect the estimates. An example of this is a prior distribution for τ_2 with a probability density function highly concentrated about a value very far from the true value of τ_2 . However, we think that if the researcher has correct information, even if it is vague, it would be worthwhile using the set of estimators $\hat{\tau}_k^P$'s.

Table 2
Relative Biases and Square Roots of Relative Mean Square Errors of the Estimators of the Population Sizes

	Population I				Population II				Population III				Population IV			
\bar{p}_1	0.05		0.01		0.05		0.01		0.05		0.01		0.05		0.01	
\bar{p}_2	0.01		0.002		0.01		0.002		0.03		0.006		0.03		0.006	
	$r\beta$	$\sqrt{r\epsilon^2}$	$r\beta$	$\sqrt{r\epsilon^2}$	$r\beta$	$\sqrt{r\epsilon^2}$	$r\beta$	$\sqrt{r\epsilon^2}$	$r\beta$	$\sqrt{r\epsilon^2}$	$r\beta$	$\sqrt{r\epsilon^2}$	$r\beta$	$\sqrt{r\epsilon^2}$	$r\beta$	$\sqrt{r\epsilon^2}$
$\hat{\tau}_1$	-0.00	0.02	-0.00	0.06	-0.00	0.02	-0.01	0.09	-0.00	0.02	-0.00	0.06	-0.00	0.02	-0.01	0.09
$\hat{\tau}_2$	0.01	0.12	0.24 ^a	0.78 ^a	0.01	0.13	0.21 ^a	0.76 ^a	0.00	0.06	0.17 ^b	0.67 ^b	0.00	0.06	0.16 ^c	0.63 ^c
$\hat{\tau}$	0.01	0.07	0.13 ^a	0.43 ^a	0.01	0.07	0.12 ^a	0.42 ^a	0.00	0.02	0.05 ^b	0.19 ^b	-0.00	0.02	0.04 ^c	0.18 ^c
$\hat{\tau}_1^U$	-0.00	0.02	-0.00	0.06	-0.00	0.02	-0.01	0.09	-0.00	0.02	-0.00	0.06	-0.00	0.02	-0.01	0.09
$\hat{\tau}_2^U$	0.02	0.13	0.14 ^a	0.65 ^a	0.01	0.12	0.14 ^a	0.65 ^a	0.00	0.06	0.13	0.65	0.00	0.06	0.13	0.71
$\hat{\tau}^U$	0.01	0.07	0.08 ^a	0.36 ^a	0.01	0.07	0.08 ^a	0.36 ^a	0.00	0.02	0.03	0.19	-0.00	0.02	0.03	0.20
$\hat{\tau}_1^J$	-0.00	0.02	-0.01	0.06	-0.00	0.02	-0.01	0.09	-0.00	0.02	-0.01	0.06	-0.00	0.02	-0.01	0.09
$\hat{\tau}_2^J$	-0.00	0.12	-0.14	0.48	-0.00	0.12	-0.14	0.48	-0.00	0.06	-0.04	0.37	-0.00	0.06	-0.04	0.35
$\hat{\tau}^J$	-0.00	0.07	-0.08	0.27	-0.00	0.07	-0.08	0.27	-0.00	0.02	-0.02	0.11	-0.00	0.02	-0.02	0.12
$\hat{\tau}_1^P$	-0.00	0.02	-0.01	0.06	-0.00	0.02	-0.01	0.09	-0.00	0.02	-0.01	0.06	-0.00	0.02	-0.01	0.09
$\hat{\tau}_2^P$	0.02	0.12	0.07	0.20	0.02	0.11	0.07	0.20	0.00	0.06	0.00	0.18	0.00	0.06	0.01	0.18
$\hat{\tau}^P$	0.01	0.06	0.04	0.11	0.01	0.06	0.03	0.11	0.00	0.02	-0.00	0.07	-0.00	0.02	-0.00	0.08

Notes: $r\beta$, relative bias; $r\epsilon^2$, relative mean square error; $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$, MLEs. Superscripts U, J , and P of estimators $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$ indicate Bayesian estimators based on the prior Uniform, Jeffrey's and two stage Poisson-Gamma distributions, respectively. Results based on 10^4 trials. Superscripts a, b and c indicate results obtained by ignoring 8%, 15% and 21% of the trials. Ignored trials were those in which the corresponding estimator of τ_2 was negative or greater than 10^4 .

Table 3
Coverage Probabilities and Average Lengths of 95% Confidence Intervals

	Population I								Population II							
	$\bar{p}_1 \approx 0.05, \bar{p}_2 \approx 0.01$				$\bar{p}_1 \approx 0.01, \bar{p}_2 \approx 0.002$				$\bar{p}_1 \approx 0.05, \bar{p}_2 \approx 0.01$				$\bar{p}_1 \approx 0.01, \bar{p}_2 \approx 0.002$			
	Bootstrap		Wald		Bootstrap		Wald		Bootstrap		Wald		Bootstrap		Wald	
	cp	\bar{l}	cp	\bar{l}	cp	\bar{l}	cp	\bar{l}	cp	\bar{l}	cp	\bar{l}	cp	\bar{l}	cp	\bar{l}
$\hat{\tau}_1^M$	NC	NC	0.95	129	NC	NC	0.94	398	NC	NC	0.93	127	NC	NC	0.76	400
$\hat{\tau}_2^M$	NC	NC	0.95	1,044	NC	NC	0.90 ^a	8,181 ^a	NC	NC	0.95	1,029	NC	NC	0.90 ^a	7,764 ^a
$\hat{\tau}^M$	NC	NC	0.95	1,052	NC	NC	0.90 ^a	8,200 ^a	NC	NC	0.95	1,037	NC	NC	0.90 ^a	7,784 ^a
$\hat{\tau}_1^D$	0.95	130	0.95	129	0.92	399	0.94	404	0.97	147	0.95	137	0.96	642	0.92	657
$\hat{\tau}_2^D$	0.94	1,110	0.95	1,044	0.74	L ₁	0.90 ^a	8,181 ^a	0.94	1,129	0.95	1,029	0.74	L ₁	0.90 ^a	7,764 ^a
$\hat{\tau}^D$	0.94	1,118	0.95	1,052	0.75	L ₁	0.90 ^a	8,201 ^a	0.95	1,139	0.95	1,038	0.78	L ₁	0.90 ^a	7,819 ^a
$\hat{\tau}_1^U$	0.94	131	0.95	129	0.92	412	0.94	403	0.97	150	0.94	137	0.97	668	0.93	657
$\hat{\tau}_2^U$	0.94	1,116	0.95	1,049	0.72	L ₂	0.89 ^a	6,887 ^a	0.94	1,128	0.95	1,028	0.73	L ₂	0.89 ^a	6,738 ^a
$\hat{\tau}^U$	0.94	1,124	0.95	1,057	0.73	L ₂	0.90 ^a	6,908 ^a	0.95	1,139	0.95	1,038	0.77	L ₂	0.90 ^a	6,796 ^a
$\hat{\tau}_1^J$	0.95	131	0.95	128	0.93	412	0.94	402	0.96	151	0.95	137	0.96	666	0.92	652
$\hat{\tau}_2^J$	0.93	1,043	0.94	998	0.58	3,122	0.71	3,142	0.93	1,057	0.93	985	0.60	3,074	0.72	3,095
$\hat{\tau}^J$	0.93	1,052	0.94	1,007	0.60	3,199	0.72	3,178	0.94	1,072	0.93	995	0.68	3,276	0.73	3,188
$\hat{\tau}_1^P$	0.94	131	0.95	129	0.91	411	0.94	402	0.89	151	0.95	137	0.86	666	0.93	654
$\hat{\tau}_2^P$	0.97	997	0.95	957	1.00	1,506	0.92	1,573	0.97	1,000	0.95	943	1.00	1,510	0.92	1,577
$\hat{\tau}^P$	0.97	1,006	0.95	966	1.00	1,575	0.94	1,624	0.97	1,011	0.95	953	1.00	1,679	0.95	1,710

Notes: cp, coverage probability; \bar{l} , average length. Superscripts *M* and *D* of the MLEs $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}$ indicate model-based and design-based confidence intervals, respectively. Bootstrap confidence intervals computed on 2,000 bootstrap samples. NC, not computed. Results based on 10⁴ trials. Superscript *a* indicate results obtained by ignoring 8% of the trials. Ignored trials were those in which the corresponding estimator of τ_2 was negative or greater than 10⁴. L₁ and L₂ indicate lengths greater than 10⁹ and 10⁴, respectively.

Table 4
Relative Biases and Square Roots of Relative Mean Square Errors of Variance Estimators

	Population I								Population II							
	$\bar{p}_1 \approx 0.05, \bar{p}_2 \approx 0.01$				$\bar{p}_1 \approx 0.01, \bar{p}_2 \approx 0.002$				$\bar{p}_1 \approx 0.05, \bar{p}_2 \approx 0.01$				$\bar{p}_1 \approx 0.01, \bar{p}_2 \approx 0.002$			
	Bootstrap		Taylor		Bootstrap		Taylor		Bootstrap		Taylor		Bootstrap		Taylor	
	rβ	$\sqrt{re^2}$	rβ	$\sqrt{re^2}$	rβ	$\sqrt{re^2}$	rβ	$\sqrt{re^2}$	rβ	$\sqrt{re^2}$	rβ	$\sqrt{re^2}$	rβ	$\sqrt{re^2}$	rβ	$\sqrt{re^2}$
$\hat{\tau}_1^M$	NC	NC	0.01	0.17	NC	NC	-0.04	0.08	NC	NC	-0.20	0.31	NC	NC	-0.64	0.65
$\hat{\tau}_2^M$	NC	NC	0.01	0.49	NC	NC	1.9 ^a	5.3 ^a	NC	NC	-0.02	0.64	NC	NC	1.8 ^a	5.4 ^a
$\hat{\tau}^M$	NC	NC	0.01	0.48	NC	NC	1.9 ^a	5.3 ^a	NC	NC	-0.02	0.64	NC	NC	1.7 ^a	5.3 ^a
$\hat{\tau}_1^D$	0.03	0.19	0.01	0.17	-0.02	0.17	-0.00	0.17	0.08	0.46	-0.07	0.28	-0.05	0.40	-0.01	0.37
$\hat{\tau}_2^D$	0.16	0.62	0.01	0.49	L ₁	L ₂	1.9 ^a	5.3 ^a	0.20	1.10	-0.02	0.64	L ₂	L ₂	1.7 ^a	5.3 ^a
$\hat{\tau}^D$	0.15	0.61	0.01	0.48	L ₁	L ₂	1.9 ^a	5.3 ^a	0.20	1.10	-0.02	0.64	L ₂	L ₂	1.7 ^a	5.3 ^a
$\hat{\tau}_1^U$	0.02	0.20	-0.01	0.17	0.03	0.19	-0.01	0.17	0.14	0.51	-0.06	0.28	0.05	0.37	0.01	0.37
$\hat{\tau}_2^U$	0.13	0.62	-0.01	0.49	0.24	1.20	1.7 ^a	4.6 ^a	0.22	0.92	-0.00	0.62	0.30	1.40	1.6 ^a	6.4 ^a
$\hat{\tau}^U$	0.13	0.61	-0.01	0.48	0.24	1.20	1.6 ^a	4.5 ^a	0.23	0.91	0.01	0.61	0.30	1.40	1.6 ^a	6.2 ^a
$\hat{\tau}_1^J$	0.06	0.21	0.02	0.17	0.05	0.19	-0.01	0.17	0.12	0.50	-0.08	0.28	0.00	0.35	-0.04	0.36
$\hat{\tau}_2^J$	0.07	0.51	-0.03	0.44	-0.25	0.66	-0.11	1.40	0.13	0.69	-0.03	0.55	-0.25	0.74	-0.13	1.50
$\hat{\tau}^J$	0.06	0.50	-0.03	0.43	-0.25	0.66	-0.12	1.40	0.12	0.68	-0.03	0.53	-0.24	0.72	-0.15	1.40
$\hat{\tau}_1^P$	0.03	0.20	-0.01	0.17	0.03	0.18	-0.02	0.17	0.16	0.52	-0.05	0.28	0.05	0.37	0.01	0.37
$\hat{\tau}_2^P$	0.07	0.34	-0.02	0.35	-0.07	0.16	-0.03	0.12	0.10	0.42	-0.01	0.41	-0.06	0.17	-0.01	0.16
$\hat{\tau}^P$	0.06	0.34	-0.02	0.34	-0.05	0.14	-0.02	0.11	0.10	0.42	-0.01	0.41	-0.03	0.15	0.01	0.16

Notes: rβ, relative bias; re^2 , relative mean square error. Superscripts *M* and *D* of the MLEs $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}$ indicate model-based and design-based variance estimators, respectively. Bootstrap confidence intervals computed on 2,000 bootstrap samples. NC, not computed. Results based on 10⁴ trials. Superscript *a* indicate results obtained by ignoring 8% of the trials. Ignored trials were those in which the corresponding estimator of τ_2 was negative or greater than 10⁴. L₁ and L₂ indicate values greater than 10² and 10⁴, respectively.

Acknowledgements

This research was supported by grant UASIN-EXB-01-01 of PROMEP and grant PAFI-UAS-2002-I-MHFM-0 of UAS. We thank Eduardo Gutierrez, the associate editor and the referees for their helpful suggestions and comments.

References

- Agresti, A. (2002). *Categorical Data Analysis*. 2nd edition. New York: John Wiley & Sons, Inc.
- Booth, J.G., Butler, R.W. and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282-1289.
- Castledine, B.J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 67, 197-210.
- Darroch, J.N. (1958). The multiple-recapture census I: Estimation of a closed population. *Biometrika*, 45, 343-359.
- Davison, A.C., and Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*. New York: Cambridge University Press.
- Evans, M.A., Kim, H.-M. and O'Brien, T.E. (1996). An application of profile-likelihood based confidence interval to capture-recapture estimators. *Journal of Agricultural, Biological, and Environmental Statistics*, 1, 131-140.
- Félix-Medina, M.H., and Thompson, S.K. (2004). Combining cluster sampling and link-tracing sampling to estimate the size of hidden populations. *Journal of Official Statistics*, 20, 19-38.
- Fienberg, S.E., Johnson, M.S. and Junker, B.W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society, A*, 162, 383-405.
- Frank, O., and Snijders, T.A.B. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.
- Heckathorn, D.D. (1994). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.
- Seber, G.A.F. (1982). *The Estimation of Animal Abundance*. 2nd edition. London: Griffin.
- Snijders, T.A.B. (1992). Estimation on the basis of snowball samples: How to weight? *Bulletin de Méthodologie Sociologique*, 36, 59-70.
- Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: What and why? *Bulletin de Méthodologie Sociologique*, 36, 34-58.
- Thompson, S.K., and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26, 87-98.

On Sample Survey Designs for Consumer Price Indexes

Alan H. Dorfman, Janice Lent, Sylvia G. Leaver and Edward Wegman¹

Abstract

Survey sampling to estimate a Consumer Price Index (CPI) is quite complicated, generally requiring a combination of data from at least two surveys: one giving prices, one giving expenditure weights. Fundamentally different approaches to the sampling process—probability sampling and purposive sampling—have each been strongly advocated and are used by different countries in the collection of price data. By constructing a small “world” of purchases and prices from scanner data on cereal and then simulating various sampling and estimation techniques, we compare the results of two design and estimation approaches: the probability approach of the United States and the purposive approach of the United Kingdom. For the same amount of information collected, but given the use of different estimators, the United Kingdom’s methods appear to offer better overall accuracy in targeting a population superlative consumer price index.

Key Words: Elementary index; Probability proportional to size sampling; Purposive sampling; Scanner data; Superlative index.

1. Background

From start to finish, survey sampling for the sake of a Consumer Price Index (CPI) must rank among the most complicated of sampling enterprises. The population target is hard to pin down, the appropriate domain of items debated, the definitions of the raw ingredients—prices, quantities, items—ambiguous and subject to question. The ultimate estimator—the estimator of the all-items CPI—relies on data from at least two surveys, one giving prices, and one giving “weights.” Below the level of “composite items” (or “item strata”)—groups of items supposed homogeneous in their price movements—there is typically no cost effective way to keep sampling weights up to date. Debate therefore goes on about the proper choice among various simple alternative estimators of price change for item categories, the “elementary aggregate indexes.” The appropriate method of aggregating these price changes, using the weights, is subject also to debate.

There are two broad approaches to the sampling by which prices are collected: probability sampling and judgment sampling. The most commonly accepted approach to survey sampling in general requires injecting an element of randomness into the survey process and relying on this randomness to make inference about population characteristics of interest—probability or “design-based” sampling; see, e.g., Särndal, Swensson and Wretman (1992). This approach was not always taken for granted. Early in the 20th century, “purposive” or “representative” sampling was considered a viable, and possibly better, option. More recently, the prediction-based school of Royall has challenged design-based assumptions; see e.g., Valliant, Dorfman and Royall (2000).

In the U.S., all CPI-related surveys are carried out using complex probability sampling techniques. Around the world, most CPI’s are constructed from judgment samples, in which experts on the different item strata choose broader or narrower classes of items for which field representatives collect prices. The fundamental reason for this is the difficulty of getting all the data one needs on the plethora of items sold, and the places where they are sold, to make probability sampling feasible.

The interesting fact is that there has been very little assessment of the relative accuracy of the different approaches to sampling. Indeed it has not been clear that it is feasible to make such assessments. The underlying population price index, for even the smallest of countries, involves so many transactions on so many items in so many places as to be inaccessible. Moreover, the population of items on the market is in a constant state of flux, complicating the application of traditional population index formulas. How then can one judge the relative closeness to “truth” of different sample-based estimates? Furthermore, in most cases, not even sample information is available for a key ingredient of the population index—namely the *quantities* of items sold—so even artificially constructing a population for test purposes from sample data has not been feasible.

The relatively recent availability of *scanner* data, in the U.S. and elsewhere, presents an unprecedented opportunity for testing sampling approaches and estimators. These data include prices *and* quantities, typically on a weekly basis, of *all* the items sold in a given category within a large sample of outlets having scanner devices. Such data may be used to construct realistic populations of transactions for which the true price index is *known*. We can then use various methods to sample from this population, construct different index

1. Alan H. Dorfman, Office of Survey Methods Research, and Sylvia G. Leaver, Office of Prices and Living Conditions, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, D.C., U.S.A., 20212; Janice Lent, U.S. Bureau of Transportation Statistics, 400 7th Street, SW, Washington, D.C., U.S.A., 20590; Edward Wegman, Center for Computational Statistics, George Mason University, Fairfax, VA, U.S.A., 22030.

estimates of interest, and compare the results to the known population parameters. One such study, described by de Haan, Opperdoes and Schut (1999), seems to show that “cutoff sampling,” the sampling of the few largest (in terms of revenue generated) items in the population, outperforms two important design-based approaches: simple random sampling (*srs*), and probability proportional to size sampling (*pps*) (where the size measure is, again, revenue).

One difficulty in any study making such comparisons is the task of maintaining a “level playing field.” If one sampling method, for example, makes use of (population) information that might not actually be available in practice, while another does not, the comparison of methods is undermined. Similarly, if one method provides only one sample or very few samples, and another provides thousands, special precautions are needed in comparing the two; indeed, such a comparison might require serious qualifications. Given the complexity of the sampling and estimation methods used in price index computation, it is not surprising that these and many other difficulties complicate experiments designed to compare various methods.

Ideally, to compare the approaches, for example, of two countries, we would mimic the whole complex sampling and estimation process of each and evaluate its costs. Both processes would be allowed the same budget, and we would be able, by some predetermined and equitable measure, to evaluate each estimate’s proximity to a known target index.

This paper comprises two studies, a large primary study, and a smaller secondary, follow-up study.

The main study is described in Sections 2 through 4. Section 2 describes the construction of the target population. Section 3 describes the “US” and “UK” methodologies and outlines the simulation details. No attempt is made to assess relative costs (thus falling short of the ideal), but the competing approaches are made as equal as possible in terms of the information they use. Results, which favor the UK approach, are given in Section 4.

The follow-up study, in Section 5, attempts to disentangle the effects of different components of the two approaches, in particular sampling method and elementary index formula. Section 6 gives a final summary and discussion.

Note on the target indexes. The price index literature contains myriad formulas for calculating price change between one period and another. Different indexes are compatible with different assumptions regarding the “average” consumer’s buying behavior in response to price change. The “fixed market basket” indexes, the commonly employed Laspeyres and less used Paasche formulas, are compatible with the assumption that consumers continue to purchase the same items in the same quantities regardless of changes in relative prices. The Laspeyres index projects the

period 1 (“base period”) quantities forward to period 2 (“current period”), while the Paasche applies the period 2 quantities to period 1. The geometric index (or “Jevons” or “*geomean*”), usually weighted with base period expenditure shares, assumes that consumers adjust the quantities they purchase in such a way that the expenditure share for each item remains constant across time. The “superlative” Fisher, Törnqvist, and Walsh index formulas, which rely on quantity (or expenditure share) information for both periods, do not require these assumptions. Formulas for these indexes, with a superscript y representing the base period, $y+1$, the current period, and i the item purchased, for the indexes are given in Appendix A.

The debate on the all items target index usually comes down to choosing between the Laspeyres and one of the superlative indexes. Most countries select a Laspeyres target index, but a strong case (Diewert 1997) can be made that the proper target is a superlative index (usually the Fisher or Törnqvist), even if the formula for the estimator does not resemble one of the superlative population index formulas. Because of the form of the US elementary aggregates – geometric mean – and the fact that previous research (Dorfman, Leaver and Lent 1999) indicated that the lowest level of estimation can have a major impact, the weighted *geomean* will be included among the potential targets. Targets for a given domain are calculated based on prices and quantities of all items in the domain following the formulas in Appendix A (a single-stage aggregation of prices and quantities).

Note: These formulas are deceptively simple, requiring the notation of section 3 for their full development. Thus, in a formula such as that for the Fisher index F (which we will take as our target in the main study of sections 2–4) “ i ” represents an item i belonging to a small class c (an “ELI” or “representative item” – see section 3), where c is itself a subset of wider classes; further, the item i is sold in a particular outlet j , classified as part of a particular chain k , and located in a particular sampled geographic area, the primary sampling unit (*psu*) l . Thus, the expression for a sum \sum_i , in the case of the overall population index, is really shorthand for $\sum_{l=1}^3 \sum_{k=1}^8 \sum_{j \in (k,l)} \sum_{C=1}^4 \sum_{h \in C} \sum_{c \in h} \sum_{i \in (j,c)}$; a similar remark holds for \prod_i . In short, these are sums and products over *all* items in the population. Contractions of this full expansion will give the population indexes for the different classes C , etc.

2. The Population for the Primary Study

The data source for the present study is a scanner data set for breakfast cereal for the years 1995 through 2000 in three separate but contiguous sections of a single large

metropolitan area. The data set was purchased from the A.C. Nielsen Corporation by the U.S. Bureau of Labor Statistics for the purpose of determining the feasibility of incorporating scanner data into the U.S. CPI; see Richardson (2000).

From these data, artificial “populations” were drawn by the method described below. Thus the study encompasses an apparently narrow world, that of cereal, within a fairly restricted geographic domain. Even this restricted world, however, allows for rather discrepant price trends over the six years. Thus, although we will not be able to generalize, in any simple fashion, to global price indexes encompassing a wide heterogeneity of products, we may be able to derive important clues on the effects of different sampling methods and the behavior of particular estimators.

The six years’ worth of data available provided the opportunity for establishing fairly long-term price trends. In order to keep the data manageable and avoid the complications of seasonality, we limited ourselves to February data. For February of year y , for each item (*i.e.*, each particular combination k of brand, type, size) in a particular outlet, four weeks t of price and quantity data were combined into a single month’s price and quantity, by using the sum of quantities $q_k^{\text{Feb}, y} = \sum_{t \in \text{Feb}, y} q_k^t$ sold during the month as the quantity, and the unit value $p_k^{\text{Feb}, y} = \sum_{t \in \text{Feb}, y} q_k^t p_k^t / \sum_{t \in \text{Feb}, y} q_k^t$ as the price. Unit values computed over short periods of time (*e.g.*, a month) give perhaps the most meaningful sense of the “average” price for a particular item. The use of unit values smoothes the data and reduces it to more manageable proportions; for a discussion of circumstances under which use of unit values is appropriate or not appropriate see Balk (1999).

For the purposes of the study, the population of breakfast cereals was divided into four groups:

1. Hot Cereals (H)
2. “Sugary” cereals (S)
3. “Fruity” cereals (F)
4. “Plain” cereals (P), *i.e.*, cold cereals not falling into categories (2) and (3).

For each group, for each successive pair of years, superlative and non-superlative indexes were calculated, using item-outlet combinations available in both years. In practice, there is generally a problem with getting perfect match-ups from period to period, and finding means to deal with this by finding substitutes for original products or by other means is important; this study bypasses this particular problem.

Long range indexes (’95 to ’00) were calculated both directly and by chaining annual indexes. Additionally, indexes were calculated on the “core” items, meaning those

available in all six years. On a year-to-year basis, the core items represented between 53% and 61% of a given year’s items available for year-to-year comparison; core expenditure was from 83% to 91% of the total expenditure on all (core and non-core) items. There were 326 core items, and a total of 848 distinct items over the course of ’95 to ’00.

Values of year-to-year population indexes are represented in Figures 1 through 5. Figure 1 gives the index $\hat{I}^{y, y+1}$ values for Sugary cereals based on all items sold in stores in both y and $y+1$, for (February of) $y = 1995, \dots, 1999$ (the “all items”). Values for five indexes are shown, including the Paasche P and, as being of academic interest, a unit value index U , the ratio of quantity weighted mean prices, averaged over all item types and outlets. Figure 2 gives results of the same calculations, but limited to “core” items. Figures 1 and 2 are almost identical, and the resemblance between indexes calculated using all items and those using just the core items held for the other cereal categories as well. Figures 3 through 5 give the results for the core-based indexes for Hot, Fruity, and Plain cereals. For any given index, the figures indicate serious differences across cereal categories. The price trends of the four major groups are quite different: H increases, S decreases sharply, F decreases modestly, and P increases modestly.

Table 1 gives long range (’95 – ’00) direct indexes and chained indexes for “all items” and “core items.” (“All items” for constructing an index between two given years, are all those item/outlets with positive quantities sold, both years). Again, there is very little difference between the values for “core items” and “all items,” and sharp differences from one cereal category to another. The chained and direct results are close for the superlative indexes but tend to be discrepant for the *geomean*, Laspeyres, and Paasche. The chained and direct unit value indexes are close and in fact the latter would be identical to the chained based on the core items, except that, for convenience, the year to year indexes were based only on item-outlet combinations available for both years.

Except for some oscillation of position of the unit value index, we observe a clear ordering of index values by formula, the same across categories, which may be summarized as follows: (1) The superlative indexes differ relatively little from each other, a noteworthy result given the amount of variability in the item-outlet price relatives and quantities, due to “sales.” (2) The traditional non-superlative indexes differ wildly from each other and the superlatives, with the *geomean*, weighted by first period expenditures, running much *higher* than the superlatives, the Laspeyres still higher, and the Paasche (not surprisingly) much lower. These results suggest that, in Cereal World, not only quantity, but expenditure share, tends to drop in period 2 on

an item having a sharp increase in price in that period. (3) The unit value index is low as well, but, except in the case of Hot cereals, is better than the traditional non-superlative indexes in approximating the superlatives. (4) In the light of later developments in this paper, and at the suggestion of a referee, two non-quantity based indexes are included in this table (although not in the figures): the *dutot* index, which is

a simple ratio of average prices (RA) – see Appendix A, and an *unweighted* (that is, constant weighted) *geomean*; both are usually reserved for computing indexes at the elementary level. The results are surprising: in approximating the superlatives, they do as well as or better than the traditional, quantity based non-superlatives, about on a par with the unit value index.

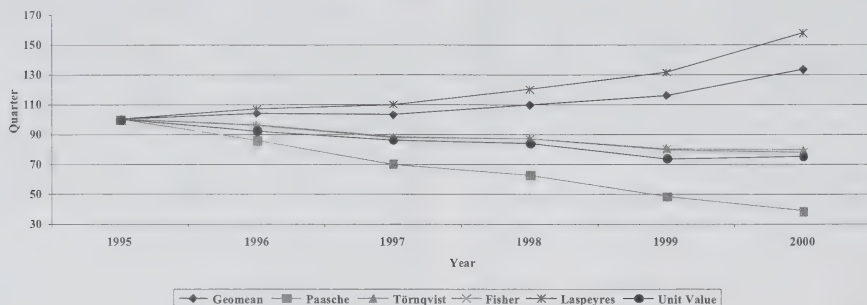


Figure 1. Annually Chained Population Target Indexes for All Sugary Cereals February-to-February Indexes, 1995 = 100.

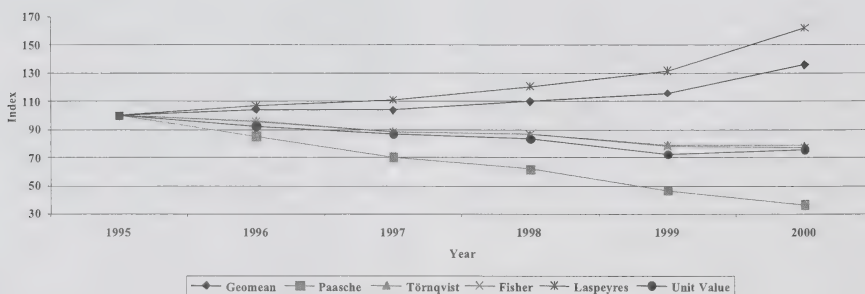


Figure 2. Annually Chained Population Target Indexes for Core Sugary Cereals February-to-February Indexes, 1995 = 100.

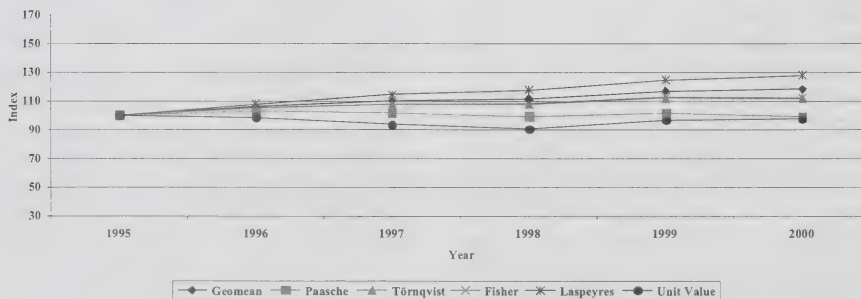


Figure 3. Annually Chained Population Target Indexes for Core Hot Cereals February-to-February Indexes, 1995 = 100.

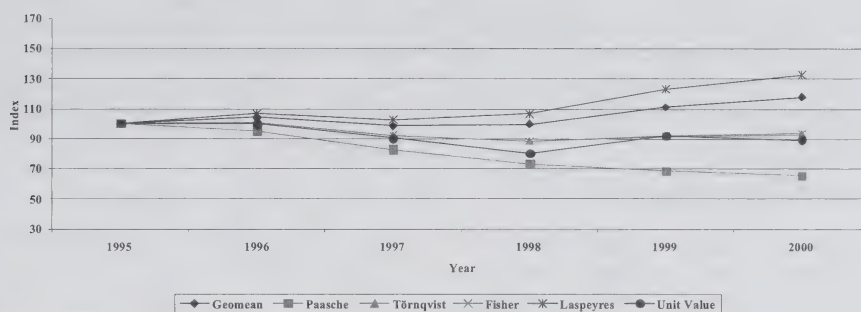


Figure 4. Annually Chained Population Target Indexes for Core Fruity Cereals February-to-February Indexes, 1995 = 100.

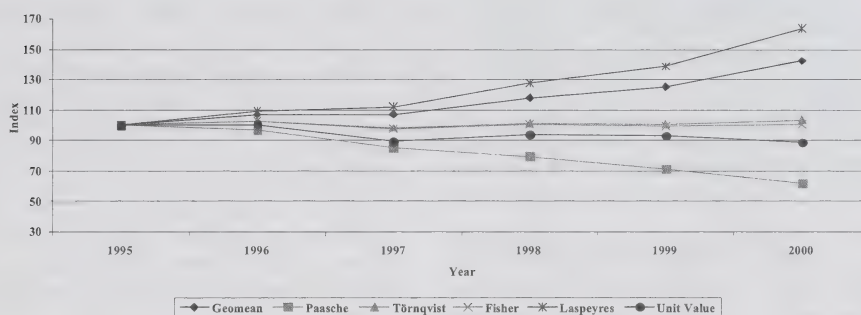


Figure 5. Annually Chained Population Target Indexes for Core Plain Cereals February-to-February Indexes, 1995 = 100.

Table 1
Direct and Chained Indexes for '95 - '00

		Geometric					
		Mean	Paasche	Törnqvist	Fisher	Laspeyres	Unit Value
Hot	Direct	1.1176	1.0253	1.0847	1.0891	1.1569	0.9576
	Chained, All Items	1.1801	0.9874	1.1159	1.1216	1.2742	0.9453
	Chained, Core Items	1.1804	0.9865	1.1160	1.1221	1.2763	0.9759
Sugary	Direct	0.8855	0.6739	0.7913	0.7898	0.9257	0.7417
	Chained All Items	1.3341	0.3825	0.7925	0.7771	1.5786	0.7506
	Chained, Core Items	1.3591	0.3661	0.7849	0.7704	1.6212	0.7585
Fruity	Direct	0.9716	0.8676	0.9319	0.9296	0.9960	0.8932
	Chained All Items	1.2202	0.6849	0.9661	0.9696	1.3728	0.9308
	Chained, Core Items	1.1808	0.6557	0.9320	0.9328	1.3269	0.8950
Plain	Direct	1.0811	0.8641	1.0045	0.9816	1.1150	0.8554
	Chained All Items	1.3969	0.6330	1.0333	1.0053	1.5965	0.8935
	Chained, Core Items	1.4234	0.6175	1.0353	1.0054	1.6370	0.8879

Table 1
Direct and Chained Indexes for '95 – '00

		Geo- metric Mean*	Paasche	Törnqvist	Fisher	Las- peyres	Unit Value	RA	Geo- metric Mean+
Hot	Direct	1.1176	1.0253	1.0847	1.0891	1.1569	0.9576	1.1192	1.0949
	Chained, All Items	1.1801	0.9874	1.1159	1.1216	1.2742	0.9453	1.1395	1.1128
	Chained, Core Items	1.1804	0.9865	1.1160	1.1221	1.2763	0.9759	1.1374	1.1151
Sugary	Direct	0.8855	0.6739	0.7913	0.7898	0.9257	0.7417	0.8817	0.8702
	Chained All Items	1.3341	0.3825	0.7925	0.7771	1.5786	0.7506	0.9124	0.9010
	Chained, Core Items	1.3591	0.3661	0.7849	0.7704	1.6212	0.7585	0.8984	0.8894
Fruity	Direct	0.9716	0.8676	0.9319	0.9296	0.9960	0.8932	0.9815	0.9726
	Chained All Items	1.2202	0.6849	0.9661	0.9696	1.3728	0.9308	1.0263	1.0165
	Chained, Core Items	1.1808	0.6557	0.9320	0.9328	1.3269	0.8950	0.9935	0.9820
Plain	Direct	1.0811	0.8641	1.0045	0.9816	1.1150	0.8554	1.0620	1.0511
	Chained All Items	1.3969	0.6330	1.0333	1.0053	1.5965	0.8935	1.0642	1.0572
	Chained, Core Items	1.4234	0.6175	1.0353	1.0054	1.6370	0.8879	1.0653	1.0571

* Weighted by base period expenditure.

+ Unweighted.

Based on this preliminary investigation, and for relative simplicity, we restricted our further investigations to the core data. To investigate the relative accuracy of probability and purposive sampling, as applied in practice to construct CPI's, we endeavored to approximate the sample designs used by the United States and the United Kingdom—representing probability based and judgment sampling respectively. In both cases we were fortunate to have detailed information on the complex survey processes, in the form of manuals, and contacts within the respective agencies. The basic idea was to repeatedly sample from a given population, for example the core transactions in the years '95 and '96. Each “run” was a composite of sampling and estimation activities carried out according to the methods of one country or the other. It should be borne in mind that our interest was in comparing the merits of *methodologies*, not in measuring the success of the US and UK in estimating their target population parameters.

The four “natural” population groups described above, referred to as “Major Groups” in the UK and “Expenditure Classes” in the US, were divided into finer sub-groups. In practice, sub-groups would be defined in terms of types of commodity. One justification for this, besides any intrinsic interest there might be in those commodities themselves, is that sub-groups so formed will tend to be homogeneous in their price trends. For purposes of this simulation study we therefore defined sub-groups as follows:

- 1) Long range price change for each of the 326 items in the core data were calculated, using unit value indexes for the items (across outlets) for '00 versus '95.
- 2) Noise was added to these indexes, items within a major group were sorted by their values of the perturbed index, and adjacent items were grouped together. The grouping of items with close long term indexes was to make the subgroups homogeneous, and the addition of noise was done so that the homogeneity would be realistically imperfect.

Table 2 gives the population item structure that was constructed, including the nomenclature in use in each of the two countries, the number of groups at each level of refinement, and the corresponding symbol for each class level used in this paper. The “Representative Item” is the lowest level at which an index is produced in the UK. This corresponds to the US's Entry Level Item (ELI), actually a collection of similar or related items. In the US, indexes are produced for categories one level up, *i.e.*, at the “Item Stratum” level, but these categories are further divided by the geographic areas in which the items are sold. Note that there are 2 or 3 Item Strata/Sections h in a Class/Major Group C , 3 ELI/Representative Items c per Item Stratum/Section h (except in one instance 2), and 10 or 11 items/varieties i in each ELI/Representative Item c . (*Note:* an actual UK class might be larger or smaller than the

corresponding US class; for example as a rule the ELI probably takes in more sorts of specific items than does the Representative Item. We had to force equivalence to ensure that the same amount of information was used in each approach. This adjustment will not affect our conclusions regarding the relative merits of the basic methods used in the two countries).

Table 2
Population Structure of “Cereal World”: Items

UK	US	Number of Groups	Symbol
Major Group	Expenditure Class	4	<i>C</i>
Section	Item Strata	10	<i>h</i>
Representative Item	Entry Level Item (ELI)	29	<i>c</i>
Variety	Item	326	<i>i</i>

In addition to the item structure, each population of transactions has a “spatial” structure, characterizing where an item was sold. This structure is summarized in Table 3. Outlets belong to chains (e.g., Safeway, Kroger), which cut across the three US geographic primary sampling units from which the cereal data were collected. (In the UK terminology, chains are called “multiples.”) Outlets in a given chain share common ownership, with the exception of “Chain 8,” which was a “catch-all” group consisting of outlets *not* belonging to a major chain (there may have been some “mini-chains”). In matching this “chain structure” to the classification of shops used in UK sampling, Chain 8 was considered a set of “independents” (the term used for independently-owned shops in the UK). Chain 4, which appeared to have the greatest homogeneity of pricing across outlets, was regarded as a “centrally collected multiple,” the term used in the UK for groups of outlets with centrally controlled pricing. Each remaining chain was a non-centrally collected multiple. The manner of collection and estimation for each of these three types is given in the description of UK methodology below.

Thus the population consists of $N^{95} \approx 20,000$ records for '95 – '96 indexes, each record representing the purchase of an item *i* within an outlet *j*. Attached to each item/outlet are its PSU/Region *l*, its chain/shop-type *k*, the outlet/shop *j*, the item/variety *i*, the ELI/representative item *c*, the item stratum/section *h*, the expenditure

class/major group *C*, and p^y, q^y, p^{y+1} , and q^{y+1} , the prices and quantities (in ounces) of the items sold in (February of) the two years in question. We used this population file (henceforth referred to simply as “the file”) to simulate all phases of the US and UK operations.

3. Sampling Methodologies Simulated

The complicated sampling procedures we used to simulate the US and UK approaches are patterned on the respective practices of these two countries. These practices change over time, and have variants even at a given point in time. Our goal was not to determine which country does better, nor to encompass all variants. Rather it was to compare two distinct modes of sampling, with the range of complexity those modes entail. The interested reader can find a description of the US construction of the CPI in the *BLS Handbook of Methods* (2005), Chapter 17. For the UK’s Retail Price Index (RPI), we relied on *The Retail Prices Index Technical Manual* (1998). A description of more current UK practice can be found in the *Consumer Price Indexes Technical Manual* (2005).

3.1 US Sampling Methodology

We first describe the US sampling methodology, which requires three surveys employing probability sampling: (1) a household survey, the Consumer Expenditure Survey (CEX), to estimate household allocation of expenditure to different categories of goods, (2) a second household survey, the Point of Purchase Survey (POPS) to estimate, within item groups, the relative amounts spent in different outlets, and (3) an outlets survey, through which individual items are selected and priced. In all three cases, sampling for the simulation is random with replacement (though the sampling employed in practice is considerably more complicated). The first two surveys are based on simple random samples, and the last is based on a probability proportional to size (*pps*) sample, where the size measures are a function of expenditures as estimated from the CEX and POPS. The sample for the third survey is a collection of items within outlet/ELI combinations.

Table 3
Population Structure of “Cereal World”: Outlets

UK	US	#	Symbol
Region	Primary Sampling Unit	3	<i>l</i>
Shop type: Independents	Chain 8		<i>k</i>
Multiples: { Central Non – central }	Chain 4		
	Chains 1 – 3; 5 – 7		
Shop	Outlet	~300	<i>j</i>

3.1.1 CEX (Household Survey)

The goal is to estimate E_{lc} , the gross household expenditure on ELI c within PSU l . We sampled using simple random sampling with replacement (*srswr*) from the file described above, within PSU, in such a manner as to get unbiased expansion estimates

$$\hat{E}_{lc}^{95} = \frac{N_l^{95}}{n_{xl}} \sum_{j \in l \cap s(xl)} \sum_{i \in c \cap s(xl)} E_{ijci}^{95}$$

where $E_{ijci}^y = q_{ijci}^y p_{ijci}^y$, N_l^{95} was the population size (number of records for *psu l* in '95 – '96), and n_{xl} was the sample size of the CEX sample $s(xl)$ in PSU l , chosen to match actual US CEX sample sizes and to achieve coefficients of variation of the estimates that approximated those achieved through the actual US CEX; the x in $s(xl)$ and n_{xl} is meant merely to differentiate the CEX from the POPS survey (which has a corresponding “ p ”; see below) or the prices survey. This “imitation CEX” was a simplified version of the actual survey. Our methodology tacitly assumed that all customers in a given outlet bought items in the same proportions; it did not allow for the inevitable measurement error that accompanies any actual expenditure survey, and (for '95 – '96) it was too current: real CEX data often predate by several years the outlet surveys for which they are used. Since, however, the “household data” collected were used in the corresponding UK methodology (see below) as well, the simplified version sufficed for the intended comparison of methodologies.

Higher level expenditures were estimated by simple addition. Thus, for example, the total expended across PSU's in a given ELI c is estimated by $\hat{E}_c^{95} = \sum_l \hat{E}_{lc}^{95}$, etc. There were 500 CEX samples taken, each producing a corresponding set of expenditure estimates.

3.1.2 POPS (Household Survey)

Here the goal is to estimate the distribution of expenditures at different outlets for particular classes of goods. These classes could be ELI's or groups of ELI's; in the present study we assume they are the ELI's. The actual US TPOPS (Telephone Point of Purchase Survey) is, as its name suggests, conducted by phone, using a sample rotation scheme with a four-year cycle. We endeavored, as we did with the CEX, to match statistical properties of our procedure to the actual TPOPS, but it turned out that to match sample sizes on our file of 20,000 would have given larger than desirable sampling fractions within PSU's. We therefore cut the sample sizes in half – our “imitation POPS” should have precision about $1/\sqrt{2}$ of the actual TPOPS. Again, this modification will not affect the conclusions of this study, because we used the identical data

in the UK construction. Samples $s(pl)$ of size n_{pl} were drawn by *srswr*, and estimation was by the expansion estimator:

$$\tilde{E}_{lci}^y = \frac{N_l^y}{n_{pl}} \sum_{i \in c \cap s(pl)} E_{ijci}^y$$

Since the POPS survey tends to be more up-to-date than the CEX, we allow y to be the base year of the index, '95 in '95 – '96, but '96 in '96 – '97, etc. There were 500 runs and sets of estimates, each to be matched with a CEX run.

3.1.3 Outlet Sampling

For each year y , selection of items from which to collect prices involves the following steps:

- (a) For each PSU l , and each of the 10 item strata h , we select 2 ELIs c by probability proportional to size with replacement sampling (*ppswr*), with size measure \hat{E}_{lc}^{95} derived from the CEX.
- (b) For each ELI c selected, we select 8 outlets j by *ppswr*, using as size measure POPS expenditure estimates \tilde{E}_{jci}^y . Thus altogether there are 160 ELI-outlet pairs per PSU, and 480 total, with a certain amount of repetition possible.
- (c) Within outlet/ELI (j, c) we “go” (as the field representative would literally go) to the outlet and “list” all items belonging to the ELI and their corresponding first period expenditures E_{ijci}^y , and, with this within-outlet frame, sample 1 item by *pps*.

For each item so selected, we record the prices p_{ijhci}^y , $y = 1, 2$. Thus we note that all aspects of the outlet sampling are *pps* with replacement, based on estimates of expenditure from one or other of the 2 household surveys or from within the selected store. Again, we performed 500 runs, each run corresponding to a single CEX/POPS run.

3.2 US Estimation

“Elementary aggregates” $\hat{I}_{ih}^{y, y+1}$, index estimates at the PSU \times Item stratum level, are the building blocks from which the CPI is constructed. In most CPI's around the world, the lowest level indexes are unweighted averages of one sort or another, as is the UK's RA estimator discussed below, and expenditure data are only used to aggregate these to higher levels. In the US, the elementary indexes are basically Horvitz-Thomson estimators relying explicitly or implicitly on expenditure estimates from both the CEX and POPS. In recent years, the US has for most item strata adopted a *geomean* formula (see Appendix A), so that estimates at this level take the form

$$\hat{I}_{lh}^{y, y+1} = \prod_{\substack{j \in l, \\ i \in c \in h, \\ (i, j) \in s}} \left(\frac{p_{ljhci}^{y+1}}{p_{ljhci}^y} \right)^{s_{ljhci}},$$

where

$$s_{ljhci} = \frac{w_{ljhci}}{\sum_{\substack{j \in l, c \in h, i \in c \\ (i, j) \in s}} w_{ljhci}},$$

with

$$w_{ljhci} = \frac{\hat{E}_{lc} \hat{E}_{lh}}{\hat{E}_{lc}} w_{ljhci},$$

$j \in l, i \in c \in h$ and $(i, j) \in s$. Note that the weights are not particular to the i^{th} item; we omit the time superscripts for brevity. They are not simply equal to the reciprocal of the number n_{lh} of sample items in lh , as sample unbiasedness considerations might lead one to expect (Balk 2003), because the sampling probabilities do not reflect exact base period expenditures on items; see the *BLS Handbook of Methods* (2005).

Then the elementary indexes are aggregated using estimated expenditures from the CEX according to a Laspeyres formula, for example

$$\hat{I}_h^{y, y+1} = \frac{\sum_l \hat{E}_{lh} \hat{I}_{lh}^{y, y+1}}{\sum_l \hat{E}_{lh}}$$

to get the index for a given item stratum h , across PSU's.

3.3 UK Sampling Methodology

The UK, like the US, combines three components in its estimation methodology: (1) a household survey, the Family Expenditure Survey (FES), to get estimates of amounts spent on different item groups, (2) a shops survey, the Annual Retailing Inquiry (ARI) to get expenditure information by section and shop type, and (3) an outlet survey of shops, to select items for pricing.

3.3.1 FES (Household Survey)

The goal is to estimate expenditures $E_{..c}$ for representative items c , and $E_{l..h}$ expenditures for region/section combinations. For purposes of this study we will assume that the data for the US's CEX and the UK's FES coincide run by run, so there are, again, 500 FES data sets. Note that the UK does not aim at the more detailed estimates $E_{l..c}$ which the US targets.

3.3.2 Annual Retailing Inquiry (Shops Survey)

The goal is to get estimates of expenditures \tilde{E}_{kh} , by section and shop type. This is considerably broader than the outlet (shop) by ELI (representative item) that the US's POPS seeks. We use the same data, for each of 500 runs, to construct the ARI estimates that we used to construct the POPS estimates for the simulated US CPI.

3.3.3 Outlet Sampling

Selecting items from which to collect prices involves the following steps:

- (a) A "judgment sample" of representative items c is selected within each section h . In the present study (only to allow for simulation), within each section, we select the two representative items with largest values of \hat{E}_{hc} . Note two differences from the corresponding step (a) of the US method: (i) selection is uniform across all regions l ; (ii) selection is not random, and, in particular, it does not allow for duplication of representative items. (Duplication can occur in the simulated US method, due to with replacement sampling of ELI's within item strata.)
- (b) The field economists select the shops in a particular locale in which to price a given representative item. Traditionally, this was *srswor*, after the field economist had constructed a frame of appropriate shops. More recently, selection has been by *pps*, where the size measure is floor space dedicated to the type of goods the representative item represents. Field economists do not draw samples of "centrally collected" items: in the case of a very large multiple, the price of an item is collected from the multiple's central office, and taken to represent the price of that item in all shops in the multiple. In the present study we proceeded as follows: for each region l and representative item c , we selected 8 shops as follows:
 - 4 from non-central multiples
(Chains 1, 2, 3, 5, 6, 7)
 - 1 from a central multiple (Chain 4)
 - 3 from independents (Chain 8)

In each case, for simplicity, we used *srs* without replacement from shops having positive expenditure for the representative item. The number of shops in the UK (8 per representative item in each region) matches the number of "outlets" in the US; there are 160 shop/representative item pairs per region, or 480 in total. Note the following differences from the U.S. methodology:

1. Information on shop type is being used for stratification (and will play a role in estimation below). This information is available in the US sample but is bypassed in favor of the *pps* methodology.
2. We are allowing the UK to have information about the presence or absence of the specific representative item *c* (equivalent to the ELI) in the list of shops before sampling, whereas the US only in effect knows of the existence of *some* ELI in the given item stratum. (This assumes a multiple ELI-to-POPS category mapping, which was typically the case until recently in US operations; the current version of ELI-to-TPOPS (telephone point of purchase survey) category mappings is 1 to 1; that is, an outlet frame is constructed for each individual ELI.)

(c) Traditionally, for each representative item *c*, within a given shop, the field economist selects that variety *i* which he/she regards as dominating its sales—a judgment sample of the most consistently purchased variety. We formalize this as follows:

1. For a given shop/representative item pair (*j*, *c*), we list all varieties *i*.
2. For each variety, we find the minimum quantity $q_i^* = \text{Min}(q_i^y, q_i^{y+1})$ over two years.
3. We sample the variety *i* with $\text{Max}\{q_i^*\}$.

This process, of course, requires more information than a field economist would have at the earlier time period (and again is not used in the US sampling described above) but may be regarded as providing a surrogate for the field economist's appraisal of the relative continuity of goods sold.

Note: It is convenient to refer to the combination of selecting an outlet by *srswor* as in (b), and an item within the shop as described in (c), as *maxminq sampling*.

3.4 UK Estimation

Elementary aggregates for the UK were calculated by a Ratio-of-Averages (RA) formula within each cross-classification cell defined by region, shop type, and representative item. This is basically an unweighted estimate, given for independent shops by

$$\hat{I}_{lkhc}^{y,y+1} = \frac{\frac{1}{n} \sum_{i \in c, j \in k} p_{ijhc}^{y+1}}{\frac{1}{n} \sum_{i \in c, j \in k} p_{ijhc}^y}.$$

In the case of multiples, a weighted version of the above formula is used with expenditures by shop type, estimated from the ARI, providing relative weights of central versus non-central multiples.

A countrywide index for representative items *c* in the sample (aggregated over shop types *k* and regions *l*) is then calculated by a Laspeyres type estimator:

$$\hat{I}_c^{y,y+1} = \sum_l \sum_k \tilde{w}_{lkhc} \hat{I}_{lkhc}^{y,y+1},$$

where $c \in h$, and \tilde{w}_{lkhc} is based on \tilde{E}_{kh}^y from the ARI and \tilde{E}_{lh}^{95} from the FES (using these time periods keeps information used the same between US and UK). Further aggregation (over representative items *c*) is done using \tilde{E}_{hc}^y , etc. from the FES.

3.5 Comparison

Table 4 gives a summary comparison of the two methodologies, US and UK, that we have been considering. The predominant feature of the US method is strict probability sampling and estimation, typically *ppswr*; that of the UK is selective sampling, taking the most important item or category as judged by expenditure or quantity sold. The methods of forming elementary aggregates are different, and the weights for aggregation in the UK are estimated at a slightly coarser level at the lower stages.

Table 5 gives a summary of what might be considered the strengths and weaknesses of the US and UK methodologies. By the advantage of "brute strength," which we attribute to the UK approach, we mean the capitalizing on a combination of two factors that often play a role in pricing and price index construction. In the first place, market leaders tend to dominate the price scene; for example, if they sharply lower or raise prices, their lesser competitors selling similar goods may think it necessary or warranted to follow suit. Secondly, even if there is variation in the price trends among similar goods, the leading sellers are likely to dominate the price index by virtue of large expenditure values, that is, because of their correspondingly large weights.

Table 4
Summary Comparison of US and UK Methodologies

	US	UK
HH Exp. Survey	\hat{E}_{lc}^{95}	$\hat{E}_{c,}^{95}, \hat{E}_{lh}^{95}$
Outlet Exp./Category	HH(POPS) \tilde{E}_{ljc}^y	Shops Survey (ARI) \tilde{E}_{kh}^y
select item categories	2 ELI's c /item stratum h /PSU l $ppswr (\hat{E}_{lc}^{95} / \hat{E}_{lh}^{95})$	2 rep. items's c /section h /Region l $largest (\hat{E}_{c,}^{95} / \hat{E}_{lh}^{95})$
select outlets	8 outlets j /ELI $c \times$ PSU l $ppswr (\tilde{E}_{ljc}^y / \hat{E}_{lc}^y)$	8 outlets j /rep. item $c \times$ Region $l - srs$ within shoptype k , $E_{ljc}^y > 0$
item within outlet/category	1 item $iljc$ $pps (E_{ljci}^y / E_{ljc}^y)$	1 variety $iljc$ $\max[\text{Min}(q_{ji}^y, q_{ji}^{y+1})]$
elementary index	$\hat{I}_{lh}^{y, y+1} = \prod_{\substack{j \in l, \\ i \in c \in h \\ (i, j) \in s}} \left(\frac{p_{ljhci}^{y+1}}{p_{ljhci}^y} \right)^{s_{ljhci}^y}$	$\hat{I}_{lkhc}^{y, y+1} = \frac{\frac{1}{n} \sum_{i \in c, j \in k} p_{ljhci}^{y+1}}{\frac{1}{n} \sum_{i \in c, j \in k} p_{ljhci}^y}$
higher aggregation	$\hat{I}_h^{y, y+1} = \frac{\sum_l \hat{E}_{lh} \hat{I}_{lh}^{y, y+1}}{\sum_l \hat{E}_{lh}}$	$\hat{I}_c = \sum_l \sum_k \hat{w}_{lkhc} \hat{I}_{lkhc}$ $\hat{w}_{lkhc} = f(\tilde{E}_{kh}, \hat{E}_{lh})$

Table 5
Comparison of US, UK Approaches

Strengths	Weaknesses
US Gather more information More use of information Satisfies classical sampling theory Gives regional (PSU) estimates Weighted estimators at lowest level More standardized operating procedure	Possible repetition in selection Ignores stratification of shops (that is, classification into chains)
UK Relies on "Brute Force" principle Stratification of outlets Shops survey in field Uses variety of sources	Patchwork of weights Inconsistent in Centralized pricing aggregation? Unweighted and seemingly arbitrary estimator at lowest level

4. Results of Primary Study

Indexes comparing '95 to '96 are given in Table 6, for the population (1) as a whole (the three areas combined), (2) broken down by classes/major groups, and (3) broken down even further into item strata/sections. Four indexes are given which might be taken as the targets of estimation. Recall the discussion on targets which concludes Section 1.

Table 7 gives corresponding means, variances, and mean square errors for US and UK estimates, where the mean square error is computed with respect to the Fisher indexes. We observe the following:

- 1) For the all-items, classes, and item strata, the US estimates appear to approximate the *geomean* G . This confirms what we have suspected from other work (Dorfman *et al.* 1999), namely that the lowest level of aggregation dominates (we used a Laspeyres formula for higher level aggregation). The fact that G lies between the Laspeyres and superlative target provides some evidence that the US switch to this method of elementary aggregation was a step in the right direction.
- 2) There appears to be no clear order relation of UK *Section* estimates to their corresponding targets; for example, the Section 11 index is higher than the target L , while the Section 12 index is lower

than the superlatives, *etc.* As we aggregate up to the Major Group and All Items levels, however, the estimates clearly begin to approximate the superlatives *F* or *T*. (Dalén (1998) noted a similar result in aggregating cut-off samples.)

- 3) If we take the Fisher as the target, *even at the section level*, the root mean square error of the UK estimator is much lower than that of the US estimator. Given the relatively restricted nature of the UK sample design, it is not surprising that the UK estimator displays lower variance, but the form of the UK estimator would not lead one to expect it to unbiasedly approximate a Fisher index. Nonetheless, our results suggest that, at least for a population of purchases such as the one used in this study, the purposive, "brute force" methods of the UK (and many other countries) work well.

Similar results were found for the succeeding pairs of years through '99 – '00. Figure 6 shows the all items year-to-year *geomean* and Fisher for five pairs of years and the means across samples of the corresponding US and UK estimators. (Note the difference in scale between Figure 6 and Figures 1 through 5). It is readily seen that the U.S. estimator tends to track the population *geomean*. The UK estimator, tracking the Fisher, tends to overestimate in the later years, although it runs much closer to the Fisher than to the population *geomean*. It should be noted that we used increasingly out-of-date expenditure data, namely the '95 data, for purposes of sampling and estimation. It is possible

that outmoded expenditure data are having a greater impact on the UK estimates than on the US estimates, perhaps by leading us to oversample expensive representative items or to focus on some group of shops that are increasingly pricey.

Results for the classes ("hot," *etc.*) were very similar for the US vis-à-vis the *geomean* and are not shown. Figure 7 shows the difference between the mean year-to-year UK estimates and the Fisher, for each of the four classes. It can be seen that the tendency to overestimate in the later years affects all four classes.

Overall, the UK estimators provide better estimates of the superlative Fisher target than do the US estimators. Table 8 gives the ratio of UK root mean square error to US root mean square error, for all five pairs of years, for all items, for groups, and for sections. There are a few anomalous places, notably in the '98 – '99 indexes where, for section 2 of "hot," and consequently for the entire class "hot," the UK estimates are appreciably worse. In general, however, the UK methods provide much better estimates. This is due in part to a tighter sampling structure (mainly because purposive/cutoff sampling is much more restrictive than random selection of the set of items which can enter the sample), yielding, not surprisingly, less variance. In part though, as well, it is due to a surprising tendency of the UK estimators to target the corresponding Fisher indexes, reducing bias. Since the UK estimators do not formally resemble the Fisher index, the reasons for their tendency to approximate it merit further study. We turn to this issue in the next section.

Table 6
Potential Target '95 – '96 Indexes

Description	<i>geomean</i>	Törnqvist	Fisher	Laspeyres
All	1.053	1.002	0.997	1.079
Classes/Major Groups				
1 – Hot	1.058	1.052	1.052	1.078
2 – Sugary	1.042	0.964	0.956	1.072
3 – Fruity	1.044	1.007	1.007	1.067
4 – Plain	1.069	1.027	1.027	1.092
Item Strata/Sections				
Hot – 11	1.043	1.044	1.044	1.057
Hot – 12	1.073	1.059	1.058	1.097
Sugary – 21	1.003	0.917	0.910	1.034
Sugary – 22	1.063	0.982	0.972	1.093
Sugary – 23	1.093	1.052	1.054	1.119
Fruity – 31	0.977	0.955	0.950	0.985
Fruity – 32	1.165	1.110	1.116	1.204
Plain – 41	1.067	1.021	1.021	1.094
Plain – 42	1.030	0.996	0.996	1.050
Plain – 43	1.104	1.063	1.062	1.125

Table 7
Simulation Results for '95 - '96 Indexes

Description	Target Index	U.S.			U.K.		
		Mean	Std. Dev.	RMSE	Mean	Std. Dev.	RMSE
All	0.997	1.057	0.016	0.062	1.002	0.011	0.012
Classes/Major Groups							
1 - Hot	1.052	1.059	0.031	0.032	1.045	0.022	0.023
2 - Sugary	0.956	1.046	0.030	0.095	0.971	0.023	0.027
3 - Fruity	1.007	1.053	0.035	0.058	0.986	0.027	0.034
4 - Plain	1.027	1.072	0.025	0.051	1.025	0.016	0.016
Item Strata/Sections							
Hot - 11	1.044	1.045	0.035	0.035	1.064	0.025	0.032
Hot - 12	1.058	1.072	0.049	0.051	1.027	0.035	0.047
Sugary - 21	0.910	1.004	0.050	0.106	0.850	0.045	0.074
Sugary - 22	0.972	1.070	0.051	0.111	1.089	0.030	0.121
Sugary - 23	1.054	1.095	0.044	0.060	1.026	0.027	0.039
Fruity - 31	0.950	0.979	0.020	0.035	0.932	0.020	0.027
Fruity - 32	1.116	1.178	0.084	0.104	1.077	0.059	0.071
Plain - 41	1.021	1.069	0.050	0.070	1.060	0.030	0.049
Plain - 42	0.996	1.033	0.035	0.051	0.987	0.031	0.032
Plain - 43	1.062	1.107	0.042	0.061	1.028	0.023	0.041

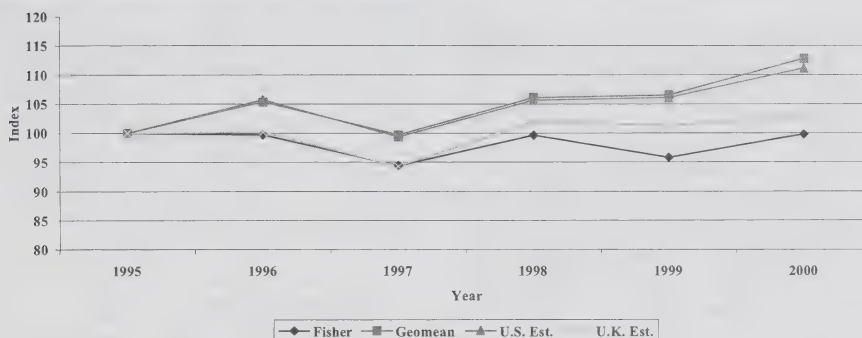


Figure 6. Index Targets and Estimates for All Cereals February-to-February Indexes and Index Estimates, 1995 = 100.

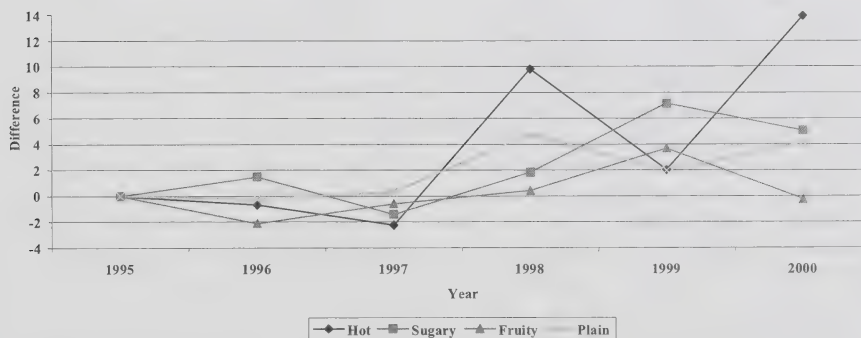


Figure 7. Differences Between U.K. Estimates and Population Fisher Indexes February-to-February Indexes and Index Estimates, 1995 = 100.

Table 8
Ratios of UK RMSE to US RMSE

Description	'95 – '96	'96 – '97	'97 – '98	'98 – '99	'99 – '00
All	0.196	0.192	0.419	0.548	0.288
Classes/Major Groups					
1 – Hot	0.713	0.517	0.483	1.437	0.589
2 – Sugary	0.286	0.336	0.314	0.522	0.282
3 – Fruity	0.595	0.508	0.308	0.501	0.405
4 – Plain	0.310	0.297	0.777	0.319	0.404
Item Strata/Sections					
Hot – 11	0.923	1.066	0.682	0.529	0.508
Hot – 12	0.920	0.850	1.169	1.860	0.842
Sugary – 21	0.702	0.392	0.421	0.595	0.330
Sugary – 22	1.092	0.426	0.380	0.341	0.365
Sugary – 23	0.650	0.455	0.448	0.925	0.851
Fruity – 31	0.778	1.059	0.637	0.581	0.618
Fruity – 32	0.683	0.809	0.314	0.457	0.356
Plain – 41	0.709	0.623	0.494	0.567	0.317
Plain – 42	0.642	0.511	1.117	1.092	1.005
Plain – 43	0.678	0.839	0.641	0.815	0.701

5. Follow-Up Study

There are four aspects in which the approaches of the UK and US differ: (1) the stratification structure, in particular, the reliance of the UK on different shops strata and, to an extent, on centralized sampling, (2) the aggregation and weighting structure, (3) the mode of sampling at different stages, and (4) the formula for elementary aggregates. This makes it difficult to disentangle the extent to which each aspect is contributing to the relative merits of US and UK index construction. In particular, as noted in the last section, it is a bit of a mystery why, especially at higher aggregations, the UK index estimator tends to target the superlative indexes.

In our follow-up study we focus on the lowest level of index construction, that is, on (3), the shop-representative item (ELI) level of sampling and on (4), the formulas for the elementary indexes. We compare the relative merits of different options, taking the within area elementary indexes as our targets. Aggregation to higher level indexes will be carried out uniformly for all alternative lower level options considered, using the true population expenditure shares. The importance of the method of construction of the elementary indexes is widely recognized; see Diewert (2004) and references; also Dorfman *et al.* (1999). The example discussed in Appendix B, with results given in Table 9, illustrates the decisive effect that the lowest level of index construction has on the index as a whole.

Thus, a likely important source of the difference in results of US and UK methodology lies in the sample estimation of the population elementary indexes. But this leaves open the question whether the differences arise because of differences in sampling method or in the

formulas used in estimation, or in both. Thus we are interested in determining: (1) how judgment sampling (in this case, cutoff sampling based on *maxming*) performs compared to probability sampling represented by *ppswr*, holding the estimator of the elementary indexes fixed, and (2) how estimators of elementary indexes compare when we keep the sampling method fixed. It will also be of some interest to determine what happens when *maxming* sampling is based on data from the base and *previous* time period, rather than the base and current period.

5.1 Sampling Methods and Estimators at the Elementary Level

To explore these questions, we carried out further simulation studies. The data were the same Cereal Data used in the primary study (successive Februarys), but limited to the Independent Shops, *Chain 8*. This was done to make the study more manageable but also because, for the other chains, the UK elementary index estimators were more complicated than the simple *dutot*. Also, it is reasonable to expect price behavior to be most heterogeneous in this chain, so that inherent differences will be clearer. Chain 8 was the largest of the chains, comprising each year about 30% of the whole population, approximately 6,000 records.

The basic structure remained the same: 3 *psu*'s, 4 major groups/expenditure classes (hot, sugary, fruity, and plain), 10 sections/item strata, and 29 representative items/*ELI*'s. For each *ELI*/representative item, 3 outlets (one item per outlet) were selected, as opposed to 10 in the primary study above. For investigating *maxming* based on previous time periods, the original 5 data sets, each using price and quantity data for a pair of years ('95/'96, '96/'97, *etc.*) were reduced to include only items that allowed "back matching";

that is, matching across three years to compare prices of items in outlets for '95/'96/'97, '96/'97/'98, etc. About 90% of the Chain 8 records allowed back matching. (In considering the results below, it is probably worth noting that the sample reduction could disproportionately impact the back matched *maxming*). We shift our attention from the Fisher index to the superlative Walsh index, due to an astute suggestion of a referee, discussed in Appendix C.

Three estimators were used for elementary indexes: the ratio of averages (RA) (the *dutot*), the unweighted *geomean* (also known as the Jevons), and the average of ratios (*AR*). In the *pps* sampling of outlets, and then in the sampling of items within outlets, the size variable (expenditure) was assumed known (rather than being estimated, as in the main study). Besides *pps* with replacement (as in the US approach), and *maxming*, we also investigated *pps* without replacement, on the suspicion it would be less variable than the with replacement version.

For each mode of sampling, within each *psu*/*ELI* combination, we took 500 samples. We calculated the mean

square error of estimates with respect to a target *ELI* – level Walsh Index. Averages of *mse* across *ELI*'s were calculated for each mode of sampling/estimation, within each *psu*.

Table 10 shows the ratio of these averages to the average *mse* for the *maxming*/*dutot* combination. For each estimator, for each *psu*, with one exception (*psu* 3, '99/'00), *maxming* leads to lower *mse*, often by an appreciable margin. Sampling *pps* without replacement is second best. Holding the method of sampling fixed (comparing rows 1, 4, 7, then 2, 5, 8, etc. in Table 10), we note that with few exceptions, the *dutot* does better than the *geomean*, which does better than *AR*. These results suggest: (1) *maxming* is better than *pps*(exp), and *pps*(exp) is better than *ppswr*(exp). (2) The *dutot* is more efficient than the *geomean*, and the *geomean* is more efficient than an average of ratios. There is a beneficial synergism between *maxming* sampling and the *dutot*. Biases and variances were also studied, and the results (not shown) tended to follow the same pattern.

Table 9
Population Indexes '95 – '96, Chain 8

Description	Laspeyres	geomean*	Fisher	Walsh	Laspeyres of Walsh Elementary
All	1.129	1.091	1.028	1.030	1.040
Classes/Major Groups					
1 – Hot	1.161	1.115	1.080	1.082	1.084
2 – Sugary	1.129	1.088	1.007	1.012	1.025
3 – Fruity	1.084	1.054	0.997	1.005	1.015
4 – Plain	1.135	1.101	1.046	1.042	1.050
Item Strata/Sections					
Hot – 11	1.157	1.117	1.088	1.089	1.090
Hot – 12	1.164	1.113	1.072	1.075	1.079
Sugary – 21	1.086	1.045	0.962	0.970	0.992
Sugary – 22	1.187	1.142	1.055	1.056	1.058
Sugary – 23	1.117	1.091	1.034	1.039	1.043
Fruity – 31	1.003	0.992	0.949	0.965	0.966
Fruity – 32	1.228	1.172	1.100	1.091	1.102
Plain – 41	1.212	1.161	1.091	1.080	1.090
Plain – 42	1.048	1.030	0.997	0.997	0.998
Plain – 43	1.136	1.107	1.048	1.046	1.056

* Weighted by base period expenditure.

Table 10
Standardized Average Relative Mean Square Error Across *ELI*'s, Reduced Populations, Chain 8

estimator/sampling method	<i>psu</i> 2					<i>psu</i> 3					<i>psu</i> 4				
	'96 – '97	'97 – '98	'98 – '99	'99 – '00	'00 – '96	'96 – '97	'97 – '98	'98 – '99	'99 – '00	'00 – '96	'96 – '97	'97 – '98	'98 – '99	'99 – '00	'00 – '96
<i>dutot</i> / <i>maxming</i> (UK)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
<i>dutot</i> / <i>ppswor</i>	1.73	1.70	1.68	1.91	1.23	1.82	1.35	2.24	1.22	1.06	1.12	0.93			
<i>dutot</i> / <i>ppswr</i>	2.13	2.10	1.91	2.14	1.42	2.10	1.46	2.67	1.45	1.23	1.36	1.07			
<i>geomean</i> / <i>maxming</i>	1.20	1.16	1.16	1.06	1.06	1.14	1.08	1.05	1.10	1.11	1.12	0.96			
<i>geomean</i> / <i>ppswor</i>	2.08	1.88	1.98	2.27	1.33	1.94	1.47	2.59	1.33	1.09	1.28	0.97			
<i>geomean</i> / <i>ppswr</i> (US)	2.49	2.29	2.18	2.53	1.58	2.23	1.58	3.09	1.59	1.30	1.52	1.12			
<i>AR</i> / <i>maxming</i>	1.42	1.32	1.31	1.14	1.24	1.03	1.30	1.05	1.11	1.20	1.21	1.07			
<i>AR</i> / <i>ppswor</i>	2.81	2.35	2.49	2.85	1.70	2.31	1.77	3.43	1.57	1.30	1.42	1.17			
<i>AR</i> / <i>ppswr</i>	3.23	2.77	2.66	3.08	2.03	2.58	1.87	3.96	1.83	1.49	1.66	1.30			
<i>dutot</i> / <i>maxming</i> , prior <i>q</i> 's	1.12	1.19	1.19	1.41	1.56	1.42	1.69	1.51	1.20	1.02	0.85	1.48			

That the *dutot* sample index can target the Walsh population index (and hence indirectly any superlative index), when consistently largest sellers are sampled is, we suggest, the result of a very simple, “brute force” mechanism: to the extent that the Walsh can be represented by a small sample of items, it is best represented by those with the consistently largest quantities, and it is these items that the *maxming* sampling scheme virtually always supplies. In Appendix C we discuss an alternative explanation for the good performance of the *maxming/dutot* combination.

Average mean square errors were also calculated for the *maxming/dutot* combination based on *previous* values of q , that is on q_i^{y-1} , q_i^y . Results are given in the last row of Table 10. There is an anticipated weakening compared to the updated *maxming/dutot*, but the results still compare favorably to the other options. We study this further in subsection 5.2.

5.2 Effect of Lagged Quantities on *maxming* Sampling

To put the results of Section 4 in perspective, we need to inquire what the effect is of using lagged q 's in *maxming*. The reason is simple: although at first sight, using base and current period quantities seems the obvious way to capture the UK's idea of persistent items, nonetheless, this involves using information (the current period quantities) which was not used in simulating US sampling. Perhaps this gives the UK methodology an unfair edge.

We therefore compared the US approach, viz. *ppswr* (with size variable being base period expenditure) and *geomean* at the elementary level, with the UK approach represented by *maxming-dutot*, but now with *maxming* based on quantities q_{j-1} and q_j . Data sets were reduced slightly to guarantee that we would have matching data for three consecutive years. Aggregation to upper level indexes used actual population expenditures for both US and UK.

Table 11 gives results for the All Cereals Indexes for chain 8, comparing biases, standard deviations, and root mean square errors with respect to the population Walsh. As expected, the results are not as good as those obtained by using current q 's. Nonetheless, with respect to all three accuracy measures (bias, standard deviation, and root mean square error), the UK *maxming/dutot* combination still does better than the US approach representing probability sampling.

For finer categories, Table 12 gives the ratios of mean square errors obtained under the UK method with lagged q 's to those obtained under the US method. Although they are generally larger than those in Table 8, they still suggest that the purposive sampling approach of the UK is better.

6. Discussion

We have presented a comparison of two fundamentally different approaches to sample design and inference for a consumer price index. The inescapable conclusion is that, in the population we studied, the “UK” approach, which involves tighter stratification and, more importantly, more restrictive judgment sampling within strata than the probability sampling of the “US” approach, does better in estimating a target superlative index.

This is shown to be the case, whichever low level price index estimator (the *dutot*, or *geomean*, or the average of ratios) is employed, although the *dutot* (ratio of averages) performed best.

The UK approach does better for two reasons: (1) its tighter sampling, restrictive of items selected (for example, see Table 13 described in Appendix C), leads, not surprisingly, to lower variance, an observation made already in de Haan *et al.* (1999), and (2) the *dutot* sample indexes target the superlative indexes under dominant market sampling, which was surprising and called forth the investigation described in Section 5. On the other hand, the US approach yielded an index estimator which could be described as unbiased, but it was unbiased for the (wrong) population *geometric* index weighted by first period expenditure. Thus it tended to run considerably higher than the target superlative index (whether Fisher, Walsh, or Törnqvist).

If sample sizes were allowed to increase, we could anticipate that the variances of both the US and UK would decrease, but the UK variance would remain lower. The bias of the US estimator for the superlative target would be unaffected by increased sample size, so that the relative mean square error of the UK approach would be increasingly lower.

In practice, of course, period 2 quantities are not available at the time of sample selection (at period 1), and as part of our follow-up study we give some measure of the partial degradation that arises from using past quantities: it is not severe enough to undo the conclusion of better UK performance. Furthermore, the field economist's judgment as to the best seller might be able to invoke data more recent than a year earlier. Thus the actual effect might be somewhere between the lagged and non-lagged versions of *maxming* which we have used. In practice, however, US field economists may often sample items within outlets based on an estimate of expenditure share that is really a smoothed average of base period and recent expenditure shares. This may attenuate the bias we have seen in our simulations, where only the base period expenditures were used for within-store sampling.

Table 11

Biases, Standard Deviations, and Root Mean Square Error (Each Multiplied by 1,000), in Estimating Population All Cereals Walsh Index, Chain 8, Based on Three Approaches to Sampling/Estimating Elementary Indexes*

	(a) Bias				
	'95-'96	'96-'97	'97-'98	'98-'99	'99-'00
<i>dutot/maxminq</i>	29	15	-13	33	2
<i>dutot/maxminq, prior q's</i>	-	46	32	82	36
<i>geomean/ppswr</i>	78	62	66	82	66
	(b) Standard Deviation				
	'95-'96	'96-'97	'97-'98	'98-'99	'99-'00
<i>dutot/maxminq</i>	16	13	11	14	12
<i>dutot/maxminq, prior q's</i>	-	14	12	15	14
<i>geomean/ppswr</i>	22	18	17	18	20
	(c) Root Mean Square Error				
	'95-'96	'96-'97	'97-'98	'98-'99	'99-'00
<i>dutot/maxminq</i>	33	20	17	36	12
<i>dutot/maxminq, prior q's</i>	-	48	34	83	39
<i>geomean/ppswr</i>	80	65	68	84	68

* At ELI/Representative Item level. To get overall index estimates, the elementary index estimates were aggregated using known population expenditures.

Table 12

Ratios of UK RMSE to US RMSE, Chain 8, Walsh Targets:
maxminq Using Lagged *q's* & *dutot* Versus *ppswr*(Expenditure) & *geomean*

Description	'96-'97	'97-'98	'98-'99	'99-'00
All	0.748	0.498	0.993	0.567
Classes/Major Groups				
1-Hot	1.539	0.495	1.280	0.765
2-Sugary	0.563	0.676	0.941	0.797
3-Fruity	0.409	0.323	0.463	0.852
4-Plain	0.915	0.560	1.164	0.359
Item Strata/Sections				
Hot-11	0.748	0.607	0.660	0.657
Hot-12	1.695	0.599	1.333	0.843
Sugary-21	0.757	0.593	1.136	0.924
Sugary-22	0.370	0.776	0.751	0.671
Sugary-23	0.479	0.785	0.796	0.508
Fruity-31	0.570	0.443	0.678	1.008
Fruity-32	0.526	0.350	0.277	0.674
Plain-41	1.167	0.509	1.395	0.397
Plain-42	0.623	0.411	0.918	0.624
Plain-43	0.919	1.171	0.668	0.560

Table 13

Items Selected by *maxminq* and *pps*($\sqrt{q_y q_{y+1}}$) in 500 Samples

'95-'96, Chain 8, <i>psu</i> 2, ELI 105											
<i>pps</i>	items selected	2889	2803	1564	2763	1558	2242	2344	2776	760	2850
	% of samples in which selected	43.2	32.2	10.4	5.4	3.87	1.53	1.33	0.87	0.8	0.4
<i>maxminq</i>	items selected	2889	2803								
	% of samples in which selected	80.87	19.13								
'95-'96, Chain 8, <i>psu</i> 3, ELI 401											
<i>pps</i>	items selected	1731	2378	2866	1742	2922	2375	2528	403	871	
	% of samples in which selected	33.27	18.8	12.8	12.73	9.47	4.6	4.27	2.8	1.27	
<i>maxminq</i>	items selected	2378	1731	2866	1742						
	% of samples in which selected	46.27	24.47	15	14.27						
'99-'00, Chain 8, <i>psu</i> 4, ELI 401											
<i>pps</i>	items selected	1731	2866	1742	2378	2922	2528	403			
	% of samples in which selected	30.07	21.93	14.3	11.07	9.53	6.8	6.27			
<i>maxminq</i>	items selected	1742	2866	2922	1731						
	% of samples in which selected	34.27	30.87	18	16.87						

It is generally accepted that the non-randomization approaches are intrinsically cheaper. For example, there are typically fewer outlets to visit, and price collection within outlets is less labor intensive. Thus, for a given budget we can expect the UK approach to be more efficient, compared to US probability sampling, than the present study suggests.

It would be salutary to expand this study to scanner data for products other than cereals. In particular, items with more volatile price movements would be of great interest. To some extent, the good behavior of *maxming/dutot* may be related to the surprising closeness of the population *dutot* to the superlatives (as seen in Table 1). How typical is such closeness, and, if it is absent, will the good sampling behavior persist?

One final *caveat*. It may be a good idea in practice to inject a dose of randomness at some stage or stages of the sampling process, and in particular be a bit cautious about centralized sampling – not for statistical reasons, but to guarantee fairness and the appearance of fairness (Reinsdorf and Triplett 2005, Section II; Royall 1976).

Acknowledgements

The opinions expressed in this paper are those of the authors and do not represent US Bureau of Labor Statistics or Bureau of Transportation Statistics policy. The authors thank David Richardson and Lyuba Rozenal for providing us with the cereal data and for timely assistance, Sonja Mapes and Scott Pinkerton for their work on the classification of cereals into types, and Mick Silver, Adrian Ball, and Dawn Camus for providing understanding and materials on the United Kingdom's RPI methods. The authors wish also to thank three referees and an associate editor for many insightful comments and for encouraging us to expand the study, and J. De Haan, M. Reinsdorf, and B. Moulton for their helpful suggestions. We especially wish to acknowledge the encouragement of the late M.P. Singh whose suggestions as Editor guided the final course this paper has taken.

Appendix A

Targets – Population Indexes

$$\text{Laspeyres } L = \frac{\sum_i q_i^y p_i^{y+1}}{\sum_i q_i^y p_i^y}$$

$$\text{Paasche } P = \frac{\sum_i q_i^{y+1} p_i^{y+1}}{\sum_i q_i^{y+1} p_i^y}$$

$$\text{Walsh } W = \frac{\sum_i \sqrt{q_i^y q_i^{y+1}} p_i^{y+1}}{\sum_i \sqrt{q_i^y q_i^{y+1}} p_i^y}$$

$$\text{Fisher } F = \left\{ \frac{\sum_i q_i^y p_i^{y+1}}{\sum_i q_i^y p_i^y} \frac{\sum_i q_i^{y+1} p_i^{y+1}}{\sum_i q_i^{y+1} p_i^y} \right\}^{1/2} = \sqrt{LP}$$

$$\text{Törnqvist } T = \prod_i \left(\frac{p_i^{y+1}}{p_i^y} \right)^{s_i^{y,y+1}},$$

where

$$s_i^{y,y+1} = \frac{1}{2} \left(\frac{p_i^y q_i^y}{\sum_i p_i^y q_i^y} + \frac{p_i^{y+1} q_i^{y+1}}{\sum_i p_i^{y+1} q_i^{y+1}} \right)$$

$$\text{Geometric Mean } G = \prod_i \left(\frac{p_i^{y+1}}{p_i^y} \right),$$

where

$$w_i = s_i^y = \left(\frac{p_i^y q_i^y}{\sum_i p_i^y q_i^y} \right)$$

or

$$w_i = 1/N$$

$$\text{Unit Value } U = \frac{\sum_i q_i^{y+1} p_i^{y+1} / \sum_i q_i^{y+1}}{\sum_i q_i^y p_i^y / \sum_i q_i^y}$$

$$\text{dutot } RA = \frac{\sum_i p_i^{y+1} / N}{\sum_i p_i^y / N} \text{ ("Ratio of Averages")}$$

$$\text{Average of Ratios } AR = \frac{\sum_i p_i^{y+1} / p_i^y}{N}$$

Appendix B

An Example Illustrating the Importance of Lowest Level Aggregation

We here present a simple example to illustrate the importance of the method used for constructing the elementary indexes. We compare population Walsh indexes to indexes resulting from aggregating elementary Walsh indexes according to a Laspeyres formula instead. The reason for focusing on the Walsh is given in Appendix C. The “pure” Walsh index is

$$W = \frac{\sum_i \sqrt{q_i^y q_i^{y+1} p_i^{y+1}}}{\sum_i \sqrt{q_i^y q_i^{y+1} p_i^y}} = \sum \tilde{s}_h W_h^{y, y+1},$$

where the $W_h^{y, y+1}$ are the h^{th} elementary Walsh indexes and

$$\tilde{s}_h = \frac{\sum_{i \in h} \sqrt{q_i^y q_i^{y+1} p_i^{y+1}}}{\sum_i \sqrt{q_i^y q_i^{y+1} p_i^y}}$$

are proper Walsh aggregation weights. To this we compare a Laspeyres aggregation of elementary Walsh indexes (“ersatz Walsh”), $L_w^{y, y+1} = \sum \sum s_h W_h^{y, y+1}$, where the s_h are standard base period weights.

The results are given in Table 9. We do see a perceptible difference between the actual population Walsh and the Laspeyres aggregate of elementary Walsh indexes: the latter tends to run slightly higher. However, these differences are on a par with the differences between them and the Fisher. They are minor compared to the gap between the *geomean* or Laspeyres indexes and the superlatives. This sort of result verifies that sound procedure at the lowest level is a key part of index construction.

Appendix C

The *maxming*/*dutot* Combination

Why does the *maxming*/*dutot* combination work so well, seeming to lead to unbiasedness for the superlative indexes?

A referee notes that *maxming* sampling bears a strong resemblance to sampling *pps* with size variable $\sqrt{q_i^y q_i^{y+1}}$; for *ppswor* ($\sqrt{q^y q^{y+1}}$), the *dutot* is approximately unbiased for a Walsh target index, and so, indirectly, for any other superlative index.

Indeed, for the expectation of the numerator of the *dutot*, under this probability sampling scheme, we have

$$\begin{aligned} E_\pi \left(\sum_{i \in s} p_i^{y+1} \right) &= E_\pi \left(\sum_{i' \in U} I_{i'} p_{i'}^{y+1} \right) \\ &= \frac{n}{\sum_i \sqrt{q_i^y q_i^{y+1}}} \sum_i \sqrt{q_i^y q_i^{y+1}} p_i^{y+1}, \end{aligned}$$

where $E_\pi()$ signifies expectation with respect to the sample design and $I_{i'}$ is a random indicator taking the values 1 or 0, as i' is in the sample or not. We get a similar expression for the denominator. The ratio of these two expected values is the Walsh. Therefore, apart from the usual (mild) ratio bias, which can be shown to be typically positive, the *dutot* does indeed target the Walsh, under this *pps* scheme.

We need to ask: do the two modes of sampling actually tend to have a sizeable overlap in what items get picked? For each run, for each *psu* l , *ELI* c , three items were selected either by *maxming* or by *ppswor* ($\sqrt{q^y q^{y+1}}$) of items within lc . Table 13 gives the percentage of times (over 500 runs) different items make it into the sample, for some arbitrarily selected representative cases. We conclude, not entirely without surprise, that: (a) *pps* sampling leads to a wider spread of items selected, (b) the items selected by *maxming* are a subset of those from *pps*, (c) there is a certain amount of correlation of “dominant items”, that is, of those items that tend most to be selected by either method. In short, *maxming* and *pps* ($\sqrt{q^y q^{y+1}}$) appear to be related, but loosely.

To get further insight into the relationship between the two sampling methods, we calculated bias and mean square error estimates, with respect to the Walsh population index, for the *dutot* index for each *ELI*, both for *maxming* and *pps* ($\sqrt{q_y q_{y+1}}$) sampling. The bias and MSE estimates were based on 500 runs for each sampling method. Summary statistics were calculated across *ELI*’s for each pair of years and each *psu*. Table 14 gives the percentage of *ELI*’s for which the *dutot* elementary indexes are positively biased for each mode of sampling. As anticipated, *pps* sampling tends to result in positive bias; we find that *maxming* is equally biased positive and negative.

Table 14
Percentage of *ELI*’s for Which the *dutot*
has Positive Bias for a Walsh Target,
for Two Sampling Schemes

	<i>pps</i> ($\sqrt{q_y q_{y+1}}$)				<i>maxming</i>			
	<i>psu</i> 2	<i>psu</i> 3	<i>psu</i> 4	<i>psu</i> 2	<i>psu</i> 3	<i>psu</i> 4		
'95-'96	75.0	86.2	75.9	64.3	61.1	61.1		
'96-'97	60.7	72.4	65.5	53.6	65.5	51.8		
'97-'98	65.5	75.9	78.6	41.4	27.6	42.9		
'98-'99	72.4	75.9	70.4	48.3	75.9	40.8		
'99-'00	89.7	72.4	75.9	48.3	20.7	44.9		

Table 15 (a) gives the percent of ELI's in which the absolute bias from using *maxming* is bigger than that from *pps* ($\sqrt{q_y q_{y+1}}$). In this regard, *pps* sampling is better. However, Table 15 (b) gives the percentage of ELI's in which *maxming* yielded a larger mean square error, and here *maxming* does better in all but two time periods/*psu*'s. We regard the mean square error criterion as the more decisive, especially given the bi-directionality of *maxming*'s biases.

Table 15

Percentage of ELI's for Which the *dutot*'s Bias and Mean Square Error for a Walsh Target is Less for Probability Proportional to Size (Size Variable = $\sqrt{q_y q_{y+1}}$) than for *maxming* Sampling

	(a) Bias of <i>pps</i> less			(b) MSE of <i>pps</i> less		
	<i>psu</i> 2	<i>psu</i> 3	<i>psu</i> 4	<i>psu</i> 2	<i>psu</i> 3	<i>psu</i> 4
'95 - '96	82.1	93.1	86.2	32.1	58.6	41.4
'96 - '97	89.2	96.6	100.0	35.7	37.9	27.6
'97 - '98	89.7	86.2	100.0	41.4	24.1	64.3
'98 - '99	89.7	82.8	92.6	41.4	37.9	40.7
'99 - '00	89.7	96.6	41.4	34.5	31.0	37.9

We conclude that the good effects of *maxming* sampling combined with the *dutot* estimator are *not* explainable in terms of approximate mimicry of *pps* sampling. They behave differently; and overall *maxming* seems to be somewhat *better* than *pps* ($\sqrt{q_y q_{y+1}}$).

We can see no alternative to explain why the *dutot* sample index should target the Walsh population index when the consistently largest sellers are sampled than that of this "brute force" mechanism: to the extent that the Walsh can be represented by a small sample of items, it is best represented by those with the consistently largest quantities, and these items are the ones the *maxming* sampling scheme supplies.

References

- Balk, B. (1999). On the use of unit values as consumer price subindices. *Proceedings of the Fourth Meeting of the International Working Group on Price Indices*, BLS, Washington, D.C.
- Balk, B. (2003). Price indexes for elementary aggregates: The sampling approach. *Proceedings of the Seventh Meeting of the International Working Group on Price Indices (Ottawa Group)*, Paris.
- BLS *Handbook of Methods* (2005). <http://stats.bls.gov/bls/descriptions.htm>.
- Consumer Price Indexes *Technical Manual* (2005). Office for National Statistics, London, http://www.statistics.gov.uk/downloads/theme_economy/CPI_Technical_Manual_2005.pdf.
- De Haan, J., Opperdoes, E. and Schut, C. (1999). Item selection in the consumer price index: Cut-off versus probability sampling. *Survey Methodology*, 25, 1, 31-41.
- Dalén, J. (1998). Studies on the comparability of consumer price indices. *International Statistical Review*, 66, 1, 83-113.
- Diewert, E. (1997). "Commentary" [on 'Alternative Strategies for Aggregating Prices in the CPI' by M.D. Shapiro and D.W. Wilcox]. Federal Reserve Bank of St. Louis Review, 79, 3, 27-37.
- Diewert, E. (2004). Index number theory: Past progress and future challenges. Presented at the SSHRC Conference on Price Index Concepts and Measurement, Vancouver, Canada, at <http://www.econ.ubc.ca/diewert/concepts.pdf>.
- Dorfman, A.H., Leaver, S.G. and Lent, J. (1999). Some observations on price index estimators. *Proceedings of the Federal Committee on Statistical Methodology Research Conference, Monday B Sessions*, 56-65.
- Reinsdorf, M., and Triplett, J.E. (2005). A review of reviews: Ninety years of professional thinking about the consumer price index. To appear, *Proceedings of the June 2004 NBER-CRIW Conference on Price Indexes*, Vancouver.
- The Retail Prices Index *Technical Manual* (1998). (Ed. M. Baxter, The Stationary Office, London, at http://www.statistics.gov.uk/downloads/theme_economy/RPI_TECHNICAL_MANUAL.pdf).
- Richardson, D.H. (2000). Scanner indexes for the CPI. *Proceedings of the Conference on Scanner Data and Price Indexes*, NBER, Cambridge, <http://www.nber.org/books/>.
- Royall, R.M. (1976). Current advances in sampling theory: Implications for human observational studies. *American Journal of Epidemiology*, 104, 463-473.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted Survey Sampling*. Springer, New York.
- Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.

An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey

Neal Thomas, Trivellore E. Raghunathan, Nathaniel Schenker,
Myron J. Katzoff and Clifford L. Johnson¹

Abstract

Researchers and policy makers often use data from nationally representative probability sample surveys. The number of topics covered by such surveys, and hence the amount of interviewing time involved, have typically increased over the years, resulting in increased costs and respondent burden. A potential solution to this problem is to carefully form subsets of the items in a survey and administer one such subset to each respondent. Designs of this type are called "split-questionnaire" designs or "matrix sampling" designs. The administration of only a subset of the survey items to each respondent in a matrix sampling design creates what can be considered missing data. Multiple imputation (Rubin 1987), a general-purpose approach developed for handling data with missing values, is appealing for the analysis of data from a matrix sample, because once the multiple imputations are created, data analysts can apply standard methods for analyzing complete data from a sample survey. This paper develops and evaluates a method for creating matrix sampling forms, each form containing a subset of items to be administered to randomly selected respondents. The method can be applied in complex settings, including situations in which skip patterns are present. Forms are created in such a way that each form includes items that are predictive of the excluded items, so that subsequent analyses based on multiple imputation can recover some of the information about the excluded items that would have been collected had there been no matrix sampling. The matrix sampling and multiple-imputation methods are evaluated using data from the National Health and Nutrition Examination Survey, one of many nationally representative probability sample surveys conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention. The study demonstrates the feasibility of the approach applied to a major national health survey with complex structure, and it provides practical advice about appropriate items to include in matrix sampling designs in future surveys.

Key Words: Missing data; Multiple imputation; Respondent burden; Split questionnaire; Sample survey.

1. Introduction

Data from sample surveys are used by researchers and policy makers in many fields. These surveys often involve nationally representative probability samples and extensive data collection based on questionnaires, and they must balance the competing goals of reasonable length and completeness in providing relevant information. The number of topics covered by such surveys, and correspondingly the amount of interviewing time involved, have typically increased over the years. The resultant increased respondent burden may be among the factors contributing to the declining response rates that have occurred. Such declining rates can result in reduced precision of survey estimates. They can also result in increased bias, if systematic differences between the nonrespondents and respondents are not accounted for in analyses of the incomplete data. Moreover, the expansion of topics covered, along with efforts to maintain high response rates, have increased the costs of conducting surveys.

One potential solution to the problem of providing the information that is needed while limiting respondent burden is to carefully form subsets of the items in a survey and administer one such subset to each respondent. Different subsets of questions (items) are administered to different subsets of respondents, so that each item is administered to at least some of the respondents. Questionnaire designs of this type are called "split-questionnaire" designs or "matrix sampling" designs, the latter name reflecting the idea that respondents (rows) and items (columns) are both "sampled" from a conceptual complete population data matrix. In many matrix sampling designs, some items (herein called "core" items) are administered to all respondents, whereas other items (herein called "split" items) are only administered to a subset of respondents. Typically, the items chosen to be core items either are especially important or are predictive of many of the split items.

The administration of only a subset of the survey items to each respondent in a matrix sampling design creates what can be considered missing data, with the missingness being

1. Neal Thomas, Datametries Research, Inc., 61 Dream Lake Drive, Madison, CT 06443, U.S.A. E-mail: snthomas99@yahoo.com; Trivellore E. Raghunathan, Department of Biostatistics and Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, U.S.A. E-mail: teraghu@umich.edu; Nathaniel Schenker, Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, U.S.A. E-mail: nschenker@cdc.gov; Myron J. Katzoff, Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, U.S.A. E-mail: mkatzoff@cdc.gov; Clifford L. Johnson, Division of Health and Nutrition Examination Surveys, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, U.S.A. E-mail: cljohnson@cdc.gov.

at random or even completely at random (Rubin 1976), since the missing data are the result of a known probability mechanism based possibly on design variables. Multiple imputation (Rubin 1987), a general-purpose approach developed for handling data with missing values, is appealing for the analysis of data from a matrix sample, because once the multiple imputations are created, data analysts can apply standard sample survey methods to the analysis of the completed data. Moreover, if the matrix sample has been designed in such a way that the items that are administered to each respondent are predictive of the items that are not administered, then the multiple-imputation approach can utilize the included items to recover information about the excluded items. We focus on multiple imputation because it is well-suited for this situation: 1) the burden of applying complex multivariate methods can be performed once by the survey organization most familiar with the design; 2) it can be implemented with existing software; and 3) it does not require novel methods for each of the numerous estimands targeted in most studies. However, alternative estimation methods to multiple imputation, both model-based and design-based, can be developed and applied to data from matrix designs.

The matrix sampling approach has been applied or explored in various settings, such as educational assessment (Sirotnik and Wellington 1977; Beaton and Zwick 1992; Zeger and Thomas 1997), health research (Wacholder, Carroll, Pee and Gail 1994; Raghunathan and Grizzle 1995; Houseman and Milton 2006), the US Census (Navarro and Griffin 1993), and business (Shoemaker 1973). Moreover, a type of matrix sampling was also used in the National Health Interview Survey of the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention prior to 1997. In that survey, chronic conditions were divided into six lists, and information about the conditions on each list was obtained from about one-sixth of the respondents (Schenker, Gentleman, Rose, Hing and Shimizu 2002). In the context of data collection for a general purpose national health survey, however, an approach to creating matrix sample designs that exploits the inherent associations among the items has not been studied.

This paper develops and evaluates a method for creating matrix sampling forms, each form containing a subset of items to be administered to randomly selected respondents. The method can be applied in complex settings, including situations in which skip patterns are present. Forms are created in such a way that each form includes items that are predictive of the excluded items, so that subsequent analyses based on multiple imputation can recover information about the excluded items that would have been collected had there been no matrix sampling. The method assumes that a training sample is available. The training sample may be

from a previous administration of a complete survey or from a pilot sample collected to support survey design. The matrix sampling method is evaluated in a study using data from the National Health and Nutrition Examination Survey (NHANES), one of many nationally representative surveys conducted by NCHS (<http://www.cdc.gov/nchs/nhanes.htm>). The NHANES, a cross-sectional survey which has been repeated several times during different time periods, obtains a large amount of data from respondents via a household questionnaire, a linked mobile-site medical examination, and laboratory analysis of biological specimens. It is of interest to examine the feasibility of matrix sampling designs for surveys such as the NHANES, which has intricate structural dependencies among its items reflected in its numerous skip patterns, along with its multiple survey components. For purposes of realism, the form-design method is applied using pilot data from the Second NHANES (NHANES II), and then the resulting design, together with multiple-imputation methods, are evaluated in a simulation study based on NHANES III data. Section 2 describes the method for designing matrix sampling forms. In Section 3, the design and results of the study based on the NHANES are described. A concluding discussion is given in Section 4.

2. Designing Matrix Sampling Forms

This section develops a method for creating matrix sampling forms, each form containing a subset of items to be administered to selected respondents.

In designing a matrix sample, it is necessary to decide which items will be core items to be included on all forms, and which items will be split items to be included on only some forms. Typically, the core items are selected based on substantive judgment as well as other considerations about the relative importance of items. Key items, for which precision of certain estimators is to be maximized, should be designated as core items, whereas less important items can be designated as split items. In addition, it is useful to select core items that are predictive of many of the split items, so that information about split items that are excluded from a form can be recovered from the core items in conjunction with the split items that are included in the form. Finally, the cost and respondent burden associated with an item are a consideration, since it can be beneficial to designate expensive and/or burdensome items as split items. The emphasis in this development is on how to allocate the split items to forms once the core items have been chosen, so it is assumed here that the core items have already been selected. However, it will be seen that the method for allocating split items uses a measure that also accounts for the usefulness of

the core items for predicting the split items. The potential to predict split items is estimated from a training sample.

It is also necessary to select a format for organizing the split items. To ensure that every pair of split items appears together on some form, so that direct estimation of all two-way associations between variables is possible, the split items are divided into blocks, and matrix sampling forms are created by putting two or more blocks of split items together (Ragunathan and Grizzle 1995). The size and number of blocks determine the length and number of forms. For example, in the study involving the NHANES to be discussed in Section 3, the split items are divided into four blocks, and each form contains two blocks (along with the core items), so there are a total of six (4 choose 2) forms. In the method developed here, the blocks are of approximately equal size, and each split item is assigned to only one block. Using blocks of the same length yields similar reduced burden for all study participants. It also yields similar precision for items of the same type. These features are not requirements for all matrix sampling designs, however. If additional precision of estimation were desired for an item, it could be included on more than one form, or it could be designated as a core item to be included on all forms.

A good matrix sampling design allocates split items to blocks in such a way that for each split item excluded from a block, there are split items included in the block that, together with the core items, are predictive of the excluded item; this facilitates the recovery of information about the excluded item during analyses of the data. The discussion below develops a method aimed at achieving this goal. The development is in two parts. First, in Section 2.1, an index is formulated for ranking how well each split item is predicted by every other split item, with predictive utility assessed as relative gain in precision conditional on the core items being included. Methods are also given for estimating the values of the index from a training sample. Second, in Section 2.2, an algorithm for assigning split items to blocks based on the index of predictive value is described.

2.1 An Index of Predictive Value

2.1.1 Preliminary Notation for Matrix Sampling Designs

Let Y denote a split item to be predicted, $X = (X_1, \dots, X_c)$ denote the core items, and Z denote a split item used to predict Y .

As mentioned above, a matrix sampling design creates what can be considered missing data. Thus, the subjects in a potential matrix sampling design can be ordered so that the n_{obs} subjects with observed values of Y are listed first, the Y -values being denoted by $Y_1, \dots, Y_{n_{\text{obs}}}$, and the n_{mis} subjects with missing Y -values follow, the Y -values

being denoted by $Y_{n_{\text{obs}}+1}, \dots, Y_{n_{\text{tot}}}$, where $n_{\text{tot}} = n_{\text{obs}} + n_{\text{mis}}$ is the total number of observations.

The expectation and variance of Y in the population targeted for matrix sampling are denoted by $E(Y)$ and $V(Y)$.

2.1.2 Simplifying Assumptions

Several assumptions are made to simplify the calculation of the index. The assumptions are used when computing the index, but not in subsequent data analyses. Each assumption could be weakened or eliminated if additional research indicates that it results in substantial degradation to the assessment of potential matrix sampling designs.

1. Each split predictor Z is considered separately when added to the core items X . If there are several items with high mutual correlation, the assignment algorithm attempts to put the items in different blocks as required for an effective matrix design. A multivariate approach based on partial correlations accounting for other split items would be anticipated to yield similar properties, and would require much more computation.
2. Each split predictor Z is assumed to be fully observed, when in practice, it will not always be available to predict Y because Z is itself a split item. Also, the occurrence of unplanned missing data (*i.e.*, missing data not created by matrix sampling) is not considered. Although these assumptions may overstate the usefulness of Z for improving estimates of $E(Y)$, such overstatement may be ameliorated via multivariate methods that utilize several variables Z . Moreover, each Z will be administered the same number of times, so any systematic bias in predictive value should be approximately the same for each split item; the primary use of the index is to order items, which is not changed by a common bias.
3. For derivation of the index values, simple random sampling is assumed for both the respondents in the matrix sample and for the training sample. Again, consistent overestimation of precision is not anticipated to substantially diminish the performance of the index.
4. It is assumed that matrix sampling produces missing data that are missing completely at random. This assumption is satisfied for all of the matrix designs considered.
5. For purposes of deriving approximations below, it is assumed that n_{tot} is large and that the ratio $n_{\text{obs}}/n_{\text{tot}}$ is fixed as n_{tot} increases. This approximation should be adequate for most estimands in national surveys.

2.1.3 Estimation Based on Multiple Imputation

The index of predictive value to be developed is based on the goal of estimating $E(Y)$ via multiple imputation applied to the matrix sample. A multiple-imputation-based estimator, $\bar{\bar{y}}$, of $E(Y)$ is approximated, under the assumption of an infinite number of imputations, as

$$\bar{\bar{y}} = \lim_{M \rightarrow \infty} M^{-1} \sum_{j=1}^M \bar{y}_j.$$

In this expression: M is the number of imputations; and \bar{y}_j is the mean from the j^{th} completed data set with imputed values $Y_{i,j}$, $i = n_{\text{obs}} + 1, \dots, n_{\text{tot}}$, and observed values $Y_{i,j} = Y_i$, $i = 1, \dots, n_{\text{obs}}$ (which do not change across completed data sets), that is,

$$\bar{y}_j = n_{\text{tot}}^{-1} \sum_{i=1}^{n_{\text{tot}}} Y_{i,j} = n_{\text{tot}}^{-1} \left(\sum_{i=1}^{n_{\text{obs}}} Y_i + \sum_{i=n_{\text{obs}}+1}^{n_{\text{tot}}} Y_{i,j} \right).$$

An estimator of the variance of $\bar{\bar{y}}$ when the imputations are created using X and Z , based on the common variance formula in Rubin (1987, Section 3.1), is

$$V_{\text{MI}} = V_{\text{comp}} + V_{\text{imp}}, \quad (1)$$

where the first term is an estimate of the variance that would be obtained with complete data, and the second term is an estimate of the variance between imputed data sets. With large samples, so that the variance of Y can be treated as known, $V_{\text{comp}} = V(Y)/n_{\text{tot}}$, and

$$V_{\text{imp}} = \lim_{M \rightarrow \infty} M^{-1} \sum_{j=1}^M (\bar{y}_j - \bar{\bar{y}})^2. \quad (2)$$

It is assumed throughout that the imputation model is compatible with the complete-data model, so the variance estimator in (2) is consistent (Rubin 1987, Section 3.6; Meng 1994).

2.1.4 Defining the Index

When matrix samples are collected, simple but potentially inefficient estimators of univariate summaries of a split item, Y , can be obtained from the observed data without any imputation (that is, using just the observed values of Y), because the subjects the missing Y -values are missing completely at random; the variance of the no-imputation estimator of $E(Y)$ is denoted by $V_{\text{NI}} = V(Y)/n_{\text{obs}}$.

The proposed index is the proportion of the difference between V_{NI} and V_{comp} that is recovered by the multiple-imputation estimator, which incorporates the information contained in X and Z :

$$I(Y|X, Z) = \frac{V_{\text{NI}} - V_{\text{MI}}}{V_{\text{NI}} - V_{\text{comp}}}. \quad (3)$$

The index $I(Y|X, Z)$ takes the value 1 when X and Z perfectly predict the omitted values of Y (so that $V_{\text{MI}} = V_{\text{comp}}$), and it takes the value 0 when X and Z do not predict the omitted values of Y at all, so that the multiple-imputation estimator is not an improvement over the no-imputation estimator (*i.e.*, $V_{\text{MI}} = V_{\text{NI}}$).

The index can be used to assess the potential contribution of each split item Z to the estimation of the mean of every other split item Y . A desirable matrix sampling design ensures that for each split item Y that is excluded from a block, there are other split items Z included in the block with high index values for predicting Y , so that information about Y can be recovered during analyses of data from the matrix sample.

Note:

1. The variances V_{NI} , V_{comp} , and V_{imp} are proportional to n_{tot}^{-1} , so $I(Y|X, Z)$ is independent of n_{tot} .
2. If the core items X are highly predictive of Y , the index will not differentiate much between the remaining split items Z ; but in this situation, the selection of appropriate Z for predicting Y is less important, since Y is already predicted well by X .

2.1.5 Approximating V_{imp}

To facilitate the computation of the index $I(Y|X, Z)$, it is useful to approximate the variance V_{imp} . The approximation developed here refers to a specific matrix sampling design, presuming that one has been chosen.

Assume that the distribution of Y given (X, Z) follows a generalized linear model with a link function μ that depends on unknown parameters β ,

$$E(Y|X, Z) = \mu((X^T, Z)\beta),$$

where the link function is equal to the identity for continuous Y , $\mu(Y) = Y$, and the logistic function for binary Y , $\mu(Y) = \text{logit}^{-1}(Y)$. For continuous Y , a constant residual variance, σ^2 , is also assumed. Although they are not developed here, extensions of these models and methods can be developed for categorical and ordered categorical variables. The individual categories can be represented by binary variables, or summaries can be formed when there are numerous categories.

Schafer and Schenker (2000) derived an approximation to the variance between imputed data sets, that is, V_{imp} , when the estimate computed from each completed data set is a smooth function g of the means of the variables involved. (In the current development, g is the identity.) Their approximation, which is based on first-order Taylor series expansions of g and μ and large-sample results

from the theory of sample surveys (e.g., Wolter 1985, Chapter 6), will be used here.

Maximum likelihood (or quasi-likelihood, McCullagh and Nelder 1989) estimation of β based on the n_{obs} subjects with observed values of Y yields an estimator, $\hat{\beta}$, with variance-covariance matrix $V_{\text{obs}}(\hat{\beta})$ (recall simplifying assumption 4 of Section 2.1.2). Set $\bar{\mu}_{\text{mis}}(\hat{\beta}) \equiv n_{\text{mis}}^{-1} \sum_{i=n_{\text{obs}}+1}^{n_{\text{tot}}} \mu((X_i^T, Z_i)\hat{\beta})$, and denote its derivative with respect to the j^{th} component of β evaluated at $\hat{\beta}$ by $\bar{\mu}'_{\text{mis},j}(\hat{\beta})$, $j = 1, \dots, (c+1)$. The derivative has the form

$$\bar{\mu}'_{\text{mis},j}(\hat{\beta}) = n_{\text{mis}}^{-1} \sum_{i=n_{\text{obs}}+1}^{n_{\text{tot}}} X_{ij} f(\hat{\beta}, X_i, Z_i) \quad j=1, \dots, c$$

and

$$\bar{\mu}'_{\text{mis},c+1}(\hat{\beta}) = n_{\text{mis}}^{-1} \sum_{i=n_{\text{obs}}+1}^{n_{\text{tot}}} Z_i f(\hat{\beta}, X_i, Z_i),$$

where $f(\hat{\beta}, X_i, Z_i) = \mu((X_i^T, Z_i)\hat{\beta})[1 - \mu((X_i^T, Z_i)\hat{\beta})]$ when Y is binary. When Y is continuous, $f(\hat{\beta}, X_i, Z_i) = 1$, which implies that the derivatives $\bar{\mu}'_{\text{mis},j}(\hat{\beta})$ are equal to the means of the core items X and the split item Z .

Now let $\bar{\mu}'_{\text{mis}}(\hat{\beta})$ denote the vector of derivatives and P_{mis} denote the proportion of subjects with missing Y . Applying equation (10) of Schafer and Schenker (2000), with their function g equal to the identity, and their general parameter θ equal to β , yields

$$V_{\text{imp}} \approx P_{\text{mis}}^2 \left[n_{\text{mis}}^{-1} \sum_{i=n_{\text{obs}}+1}^{n_{\text{tot}}} \mu((X_i^T, Z_i)\hat{\beta}) [1 - \mu((X_i^T, Z_i)\hat{\beta})] + (\bar{\mu}'_{\text{mis}}(\hat{\beta}))^T V_{\text{obs}}(\hat{\beta}) \bar{\mu}'_{\text{mis}}(\hat{\beta}) \right] \quad (4)$$

when Y is binary, and

$$V_{\text{imp}} \approx P_{\text{mis}}^2 [\sigma^2 / n_{\text{mis}} + (\bar{\mu}'_{\text{mis}}(\hat{\beta}))^T V_{\text{obs}}(\hat{\beta}) \bar{\mu}'_{\text{mis}}(\hat{\beta})] \quad (5)$$

when Y is continuous.

2.1.6 Estimating the Index of Predictive Value from a Training Sample

Because the planned missing data in our matrix sampling designs are assumed to be missing completely at random, sample moments and other parameter estimates from a training sample can be used to estimate the corresponding moments and parameters in both the subsamples with observed and missing values of Y , under the assumption that the training sample is drawn from the same target population. The moments and parameters include: $V(Y)$; the residual variance σ^2 , which can be estimated by $\hat{\sigma}_{\text{tr}}^2$; the estimated residual variance from the regression fitted to

the training sample; the regression coefficients, with estimates $\hat{\beta}_{\text{tr}}$ from the training sample; and the variance-covariance matrix of the regression coefficients, which can be approximated by rescaling the estimate $V_{\text{tr}}(\hat{\beta}_{\text{tr}})$ from the training sample to obtain $V_{\text{obs}}(\hat{\beta}) \approx (n_{\text{tr}}/n_{\text{obs}})V_{\text{tr}}(\hat{\beta}_{\text{tr}})$, where n_{tr} is the size of the training sample. The derivatives $\bar{\mu}'_{\text{mis}}(\hat{\beta})$ and the function involving μ in (4) are also in the form of subsample means, and thus can be estimated by the corresponding means in the training sample. Denoting the derivatives in the training sample by $\bar{\mu}'_{\text{tr}}(\hat{\beta}_{\text{tr}})$, and substituting the training sample estimators into (4) and (5), yields

$$V_{\text{imp}} \approx P_{\text{mis}}^2 \left[n_{\text{mis}}^{-1} \left\{ n_{\text{tr}}^{-1} \sum_{i=1}^{n_{\text{tr}}} \mu((X_i^T, Z_i)\hat{\beta}_{\text{tr}}) [1 - \mu((X_i^T, Z_i)\hat{\beta}_{\text{tr}})] \right\} + \frac{n_{\text{tr}}}{n_{\text{obs}}} (\bar{\mu}'_{\text{tr}}(\hat{\beta}_{\text{tr}}))^T V_{\text{tr}}(\hat{\beta}_{\text{tr}}) \bar{\mu}'_{\text{tr}}(\hat{\beta}_{\text{tr}}) \right] \quad (6)$$

for binary variables, and

$$V_{\text{imp}} = P_{\text{mis}}^2 (\hat{\sigma}_{\text{tr}}^2 / n_{\text{mis}} + \hat{\sigma}_{\text{tr}}^2 / n_{\text{obs}}), \quad (7)$$

for continuous variables, with the latter expression following from the fact that $(\bar{\mu}'_{\text{tr}}(\hat{\beta}_{\text{tr}}))^T V_{\text{tr}}(\hat{\beta}_{\text{tr}}) \bar{\mu}'_{\text{tr}}(\hat{\beta}_{\text{tr}})$ reduces to the simple form $\hat{\sigma}_{\text{tr}}^2 / n_{\text{tr}}$.

2.2 Assigning Split Items to Blocks

2.2.1 Design Criteria

Matrix sampling forms are created by allocating split items to different blocks, as described at the beginning of Section 2. Four design goals guide the assignment of items: 1) assign each split item to a single block; 2) assign an approximately equal number of items to each block; 3) assign logically linked items to the same block; and 4) assign one or more items to each block that predict the items omitted from the block. Denote the number of blocks by n_{block} ($n_{\text{block}} = 4$ in the NHANES simulation study).

A quantitative criterion for the fourth goal is specified separately for each split item Y by finding the $(n_{\text{block}} - 1)$ other split items Z with the highest predictive index values $I(Y|X, Z)$, for the potential allocation of one of them to the $(n_{\text{block}} - 1)$ blocks not containing Y . The items Z exclude those items linked to Y , which must appear with Y in a block. The $(n_{\text{block}} - 1)$ values of $I(Y|X, Z)$ for the items Z provide an upper limit on the predictive indices that could be achieved for Y . Because these optimal index values are determined separately for each split item Y , they may not be achievable for all items Y simultaneously.

To evaluate a given matrix sampling design, the highest index value $I(Y|X, Z)$ actually achieved for each of the $(n_{\text{block}} - 1)$ blocks not containing a split item Y is

determined. The average of the $(n_{\text{block}} - 1)$ differences between these indices and the corresponding optimal predictive indices for Y is computed. These average differences are then averaged across all of the split items Y to yield an overall measure for the design.

2.2.2 An Assignment Algorithm

The criteria in Section 2.2.1 require maximization over a set of integer inputs (block assignments) to a function subject to a set of linear constraints imposed by the need to create approximately equal-length blocks with some items potentially linked together. Although integer programming methods could be applied to this maximization, the following algorithm is much simpler, and it achieved nearly optimal results for the NHANES application, as demonstrated in Section 3.1.

Step 1. Randomly order the split items. The assignment of items to blocks proceeds sequentially, via repetition of steps 2 and 3 below, until all of the items have been assigned.

Step 2. Assign the next (or first) unassigned item, say $Y^{(0)}$, to the block with the fewest items. If multiple blocks are tied, assign $Y^{(0)}$ to the block with the lowest maximum predictive index $I(Y^{(0)} | X, Z)$ for $Y^{(0)}$. If a tie still remains, assign $Y^{(0)}$ to any of the eligible blocks. If there are items linked to $Y^{(0)}$, also assign them to the selected block.

Step 3. For each item assigned in Step 2 ($Y^{(0)}$ or its linked items), find the remaining unassigned item, say $Y^{(1)}$, most predictive of it. Assign $Y^{(1)}$ (and any items linked to $Y^{(1)}$) to a block other than the block selected in step 2, by following the same procedure that was used for $Y^{(0)}$ in step 2.

Experience with the NHANES data suggests moderate sensitivity of the algorithm to the initial ordering of the items (in step 1). To reduce the dependence, 1,000 designs were generated with randomly selected orderings, and the one yielding the best overall measure of predictive value (as defined at the end of Section 2.2.1) was selected.

3. A Study Using Data from the NHANES

To assess the feasibility of a matrix sampling design for a survey like the NHANES, an evaluation study was conducted. First, NHANES II (*i.e.*, the second NHANES) was used to create a matrix sampling design via the method described in Section 2. This simulates the realistic situation in which data from a previous survey are used in designing the questionnaire for a new survey. The design so developed was then applied to several simulated samples created from NHANES III data. The subjects in NHANES III with complete data on a selected set of variables were treated as a large finite population. One hundred samples were drawn

from the NHANES III finite population using a stratified two-stage sample design with unequal probabilities of selection. The complete data are available for each simulated sample, providing a “gold standard.” The matrix sampling design was then imposed on each sample, and the missing values due to matrix sampling were multiply imputed. Several analyses were conducted using the matrix samples without imputation, the multiply imputed matrix samples, and the samples of complete data (*i.e.*, the gold standard). Results summarized across the simulated samples yield estimates of the repeated sampling properties of the different methods.

The matrix sampling design created using NHANES II data is summarized in Section 3.1. The design of the simulation study using NHANES III data is described in Section 3.2. Results of the study are presented in Section 3.3. Some limitations of the study that are not discussed in Sections 3.1–3.3 are covered in Section 3.4.

3.1 A Matrix Sampling Design Based on Training Data from NHANES II

Given the time that would have been required to extract and analyze all of the NHANES III variables, only a subset were included in the study to keep it manageable, although the software utilized in the study could be applied with many more variables. The variables in the study include items representing many of the topics included in the survey and were selected in consultation with substantive experts. The data types include binary and continuous variables representing survey questions and laboratory measurements. One pair of items forming a skip pattern was included: “Have you smoked 100+ cigarettes?” followed by “Do you smoke now?” The algorithm for assigning split items to blocks, described in Section 2.2, forced these items to be in the same block.

Table 1 gives brief descriptions of the variables included. Variables that appeared in NHANES III but not in NHANES II (again, a realistic situation) have asterisks next to their names.

As mentioned earlier, the matrix sampling design was constructed with four blocks. Each block contained all of the core items. In addition, the split items that appeared in NHANES II were allocated to the blocks by applying the methods developed in Section 2 to data from NHANES II. (In estimating the necessary indices, missing values in the NHANES II data were handled by analyzing only the complete cases.) The split items that did not appear in NHANES II were randomly divided and assigned to the blocks to keep the block lengths approximately equal. The “Type” column of Table 1 identifies the core and split variables and indicates the block assignments for the split variables.

For each split item that appeared in NHANES II, Table 2 displays the following: the block to which the item was assigned (“Block”); the three highest predictive indices for other split items as predictors of the item in question (“Optimal”); and the highest index values actually achieved by the selected design in the three blocks not containing the item in question (“Achieved”). The index values are sorted from low to high for each item in question, so the columns in the table containing index values do not correspond to specific items or blocks. Table 2 shows that the selected design is nearly optimal for the criteria of Section 2.2.1. For

example, the average difference between the optimal predictive indices and the corresponding indices actually achieved is only 0.002.

The column of Table 2 labeled “Low” under “Achieved” provides lower bounds on the anticipated improvement in estimators of univariate means for the split items. Nineteen of the twenty-one predictive indices in this column are less than 0.20, suggesting relatively low efficiency for multiple-imputation estimators in this matrix sampling design. For more discussion of this issue, see Sections 3.3 and 4.

Table 1
Variables from NHANES III that were Included in the Evaluation.
Items Marked with Asterisks did not Appear in NHANES II

Variable Name	Description of the Variable	Type
BMPBMI	Body Mass Index	Core
CHP*	Serum Cholesterol (MG/DL)	Core
DMARETHN	Race-Ethnicity	Core
DMPCREGN	Census Region, Weighting(Texas in South)	Core
DMPMETRO	Rural/Urban Code Based on Usda Code	Core
GHP*	Glycated Hemoglobin: (%)	Core
HAB1	Is Health in General Excellent, ..., Poor	Core
HAB2*	Go to Particular Place for Health Care	Core
HAB5*	Past 12 Months, # Times Saw Doctor	Core
HAC1C	Doctor Told: Congestive Heart Failure	Core
HAC1L*	Doctor Ever Told you Had: Lupus	Core
HAC1M	Doctor Ever Told you Had: Gout	Core
HAD1	Ever Been Told you Have Sugar/Diabetes	Core
HAD10	Are you Now Taking Diabetes Pills	Core
HA3	Told 2+ Times you Had Hypertension/HBP	Core
HAF10	Doctor Ever Told you Had a Heart Attack	Core
HAF26	Severe Dizziness for More Than 5 Minutes	Core
HAL1	Cough Most Days, 3+ Consecutive mo in YR	Core
HAL6	Had Wheezing, Whistle in Chest Past 12 MO	Core
HAL14E	Symptoms Brought on by: Pollen	Core
HAZMNK1R	Average K1 BP from Household and MEC	Core
HAZMNK5R	Average K5 BP from Household and MEC	Core
HFA12	Marital Status	Core
HFA8R	Highest Grade or YR of School Completed	Core
HSAGEIR	Age at Interview (Screener) – Qty	Core
HSSEX	Sex	Core
IIP	Serum Insulin (UU/ML)	Core
G1P	Plasma Glucose (MG/DL)	Split – 1
HAC1J	Doctor Ever Told you Had: Goiter	Split – 1
HAC1N*	Doctor Ever Told you Had: Skin Cancer	Split – 1
HAC1O	Doctor Ever Told you Had: Other Cancer	Split – 1
HAF14*	Get Pain in Either Leg While Walking	Split – 1
HAL11A	Stuffy, Itchy, or Runny Nose, Past 12 MO	Split – 1
BMPWHR*	Waist to Hip Ratio	Split – 2
HAC1E	Doctor Ever Told you Had: Asthma	Split – 2
HAC1K	Doctor Ever Told you Had: Thyroid Disease	Split – 2
HAF24	Numbness etc, 1 Side Face/Body for > 5 Min	Split – 2
HAL11B	Watery, Itchy Eyes in Past 12 Months	Split – 2
HAL19A*	In Past 12 Months Had: Cold or Flu	Split – 2
HAL19C*	In Past 12 Months Had: Pneumonia	Split – 2
HAT28	Active Compared with Men/Women your Age	Split – 2
PBP	Lead (UG/DL)	Split – 2
SPPFVC*	FVC, Largest Value (ML)	Split – 2

Table 1 (Continued)

Variables from NHANES III that were Included in the Evaluation.
Items Marked with Asterisks did not Appear in NHANES II

Variable Name	Description of the Variable	Type
FEP	Serum Iron (UG/DL)	Split – 3
HAF1	Ever Had Any Pain or Discomfort in Chest	Split – 3
HAF23	Weak/Paralysis on Face, Arm, Leg For > 5 Min	Split – 3
HAL19B	In Past 12 MO Had: Sinusitis/Sinus Prob	Split – 3
HAR1	Have you Smoked 100+ Cigarettes In Life	Split – 3
HAR3	Do you Smoke Cigarettes Now	Split – 3
SPPPEAK*	Peak Expiratory Flow	Split – 3
BDPTOBMD*	Bone Mineral Density Total Region-GM/CM SQ	Split – 4
HAB4	Past 12 MOS, # Times Stayed in Hospital	Split – 4
HAC1D	Doctor Ever Told you Had: Stroke	Split – 4
HAC1F	Doctor Ever Told Had: Chronic Bronchitis	Split – 4
HAC1H	Doctor Ever Told you Had: Hay Fever	Split – 4
HAC1I	Doctor Ever Told you Had: Cataracts	Split – 4
HAE6*	Ever Had Blood Cholesterol Checked	Split – 4
HAM11*	Consider Self Over/Under/Right Weight	Split – 4
HAE7*	Doctor Told Blood Cholesterol Level High	Split – 4

Table 2

Indices of Predictive Value Based on NHANES II Data for the Split Items in the Matrix Sampling Design

Item	Block	Low	Optimal Medium	High	Low	Achieved Medium	High
HAC1J(GOITER)	1	0.04	0.04	0.15	0.04	0.04	0.15
HAC1O(OTHER CANCER)	1	0.05	0.06	0.13	0.05	0.06	0.13
HAL11A(NASAL SYMPTOMS)	1	0.17	0.27	0.29	0.17	0.27	0.29
G1P(PLASMA GLUCOSE)	1	0.26	0.30	0.43	0.26	0.30	0.43
HAC1E(ASTHMA)	2	0.09	0.10	0.13	0.08	0.09	0.13
HAC1K(THYROID DISEASE)	2	0.07	0.07	0.15	0.07	0.07	0.15
HAF24(NUMBNESS)	2	0.12	0.12	0.12	0.11	0.12	0.12
HAL11B(WATERY EYES)	2	0.14	0.15	0.25	0.14	0.15	0.25
HAT28(ACTIVE FOR AGE)	2	0.12	0.13	0.16	0.11	0.13	0.16
PBP(LEAD (UG/DL))	2	0.19	0.20	0.21	0.18	0.20	0.21
HAF1(PAIN IN CHEST)	3	0.25	0.29	0.29	0.23	0.25	0.29
HAF23(WEEK/PARALYSIS)	3	0.08	0.12	0.12	0.08	0.12	0.12
HAL19B(SINUSITIS/SINUS)	3	0.07	0.12	0.21	0.07	0.12	0.21
HAR1(100+ CIGARETTES)	3	0.13	0.14	0.14	0.13	0.14	0.14
HAR3(SMOKE NOW)	3	0.10	0.11	0.12	0.10	0.11	0.12
FEP(SERUM IRON)	3	0.05	0.05	0.08	0.05	0.05	0.08
HAB4(# HOSP STAYS)	4	0.07	0.11	0.19	0.07	0.11	0.19
HAC1D(STROKE)	4	0.19	0.20	0.24	0.18	0.20	0.24
HAC1F(BRONCHITIS)	4	0.10	0.12	0.12	0.10	0.10	0.12
HAC1H(HAY FEVER)	4	0.07	0.07	0.09	0.04	0.07	0.09
HAC1I(CATARACTS)	4	0.08	0.09	0.12	0.08	0.09	0.12

Note: Optimal predictive indices are determined for each item separately and may not be achievable for all items simultaneously.

3.2 Design of the Simulation Study Based on NHANES III Data

3.2.1 Population and Sample Design

The matrix sampling design and multiple-imputation analysis could be applied to the entire NHANES III sample. Although this would be informative, a study based on a single data set would not allow the assessment of repeated-sampling statistical properties of the methods studied. Therefore, the 11,759 subjects from the NHANES III

survey who had complete data on the variables listed in Table 1 were treated as a finite population, and repeated samples were drawn from this population. In selecting samples, a complex sample design was used instead of simple random sampling to create a more realistic simulation study. To achieve this objective, three design variables were added to the finite population: (1) simulation stratum; (2) simulation cluster; and (3) simulation sample weight (here, the modifier "simulation" is used to distinguish these quantities from the original NHANES III design variables).

1. **Simulation strata:** The NHANES III public-use sample has 49 strata with two clusters per stratum. The strategy for the simulation study was to create a smaller number of strata with a larger number of clusters within the strata, to ensure sufficient sample-to-sample variation between the simulated samples. The 49 original strata were collapsed into 20 simulation strata as follows. Each of the 49 original strata were classified into one of eight categories formed from the cross-classification of census region (4 levels) and rural/urban status based on the United States Department of Agriculture code (2 levels). Within each of these eight categories, a cluster analysis was performed using the stratum-level proportions of non-Whites to select the original strata to combine. Combining the original strata created two or three simulation strata within each of the eight categories, yielding a total of 20 simulation strata. This method of creating larger strata also increased the racial heterogeneity between the resulting simulation strata, which increases the importance of weighting in the analyses.
2. **Simulation clusters:** The NHANES III public-use sample has 98 clusters, with two clusters in each of the original 49 strata. After the 49 original strata were collapsed into 20 simulation strata, the original clusters were subdivided based on another cluster analysis using systolic and diastolic blood pressure readings and body mass index (BMI). Subjects with similar values were grouped together to create a setting with intraclass correlation for these three variables within each simulation cluster. The number of simulation clusters per simulation stratum ranged from 3 to 25, and the number of subjects per simulation cluster ranged from 30 to 98.
3. **Simulation sampling weights:** The simulation sampling weights were determined by the following two-stage sample design. First, from each simulation stratum, two simulation clusters were drawn via simple random sampling without replacement. Because there were unequal numbers of simulation clusters across the 20 simulation strata, the simulation sampling weight corresponding to this stage was $w_{1h} = A_h / 2$, $h = 1, 2, \dots, 20$, where A_h is the number of simulation clusters in simulation stratum h . Second, from each selected simulation cluster, 30 subjects were drawn at random without replacement with varying probabilities of selection. If the cluster size was 30, then all subjects were included in the

sample. For clusters with more than 30 subjects, the first-draw selection probabilities were computed by normalizing the reciprocals of the original weights from the NHANES III public-use sample to sum to 1 within each simulation cluster, with the normalized reciprocal for each subject used as the selection probability for that subject. The first-draw selection probabilities within simulation clusters ranged from 0.0003 to 0.2756.

Let i index sampled subjects within a simulation cluster, c denote sampled clusters within a simulation stratum, and h denote simulation strata as above, $i = 1, 2, \dots, 30$, $c = 1, 2$, $h = 1, 2, \dots, 20$. If the size of cluster c in stratum h was 30, then the second-stage simulation weight for subject i in cluster c was $w_{2ich} = 1$. If the size of cluster c in stratum h was greater than 30, then the second-stage simulation weight for subject i in cluster c was $w_{2ich} \propto \pi_{ich}^{-1}$, where π_{ich} denotes the first-draw selection probability for subject i . The final simulation sampling weight for each sampled subject was $w_{ich} = w_{1h} \times w_{2ich}$, $i = 1, 2, \dots, 30$, $c = 1, 2$, $h = 1, 2, \dots, 20$.

The design effects for estimating population means averaged approximately 2.1 in this simulation study. The complex sample design features in the study are informative in the sense that ignoring the design features in analyses of data may result in biased estimates and underestimation of sampling variances. This is due in particular to the use of data on race, blood pressure, and BMI in the simulation sampling design, and the well-documented connection between race/ethnicity and blood pressure or BMI.

3.2.2 Simulating Matrix Samples

One hundred independent probability samples were drawn from the finite population. Each simulated sample included 1,200 subjects (20 simulation strata, 2 simulation clusters per simulation stratum, 30 subjects per simulation cluster).

Matrix sampling was overlayed on each simulated sample by assigning each of the 1,200 subjects randomly to one of the six forms containing the core items and one of the block pairs (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), or (3, 4). The random assignment was carried out such that 200 subjects were assigned to each form. Thus, for each matrix sample, the core items were available for all 1,200 sampled subjects, whereas each split item was available for 600 sampled subjects.

3.2.3 Estimation Methods Compared

Point estimates from each sample in the simulation study were obtained using three methods: analyzing the complete

data for a gold standard; analyzing the matrix sampled data with no imputation; and applying multiple imputation to fill in the missing values caused by matrix sampling, followed by multiple-imputation analyses. For the complete-data and no-imputation analyses, the point estimates were weighted. For the multiple-imputation analyses, the same weights were used in calculating the point estimate from each of the multiple completed data sets, and then the usual averaging of the multiple point estimates was carried out (Rubin and Schenker 1986; Rubin 1987, Section 3.1).

Multiple imputation of the missing split items was carried out using the sequential regression approach (Kennickell 1991; Oudshoorn, Van Buuren and Van Rijkevorsel 1999; Raghunathan, Lepkowski, Van Hoewyk and Solenberger 2001), as implemented by the software package IVEware (<http://www.isr.umich.edu/src/smp/ive>). Five sets of imputations were created by independently applying the sequential regression approach five times, with ten iterations of the sequential regression algorithm for each set of imputations. The number of imputations is based on theory and experience showing that five imputations is usually adequate, especially if the fraction of missing information is not large (Rubin 1996). With missing-data rates for the split items of 50%, the fraction of missing information, which is roughly $1 - V_{\text{comp}} / V_{\text{imp}}$, is expected to be at most 50%, as is borne out in the simulation results. Rubin (1987, Table 4.1) gave the large-sample relative efficiency of five imputations relative to an infinite number of imputations as 90% when there is 50% missing information. A larger number of imputations would increase precision for estimating the between-imputation variance (V_{imp}) and the fraction of missing information.

To account for the complex simulation sample design, main effects were included in the imputation model for simulation stratum and simulation cluster nested within simulation stratum. The logarithm of the simulation sampling weight was also included as a predictor in the imputation model, along with the core and split items.

3.3 Results of the Simulation Study

To evaluate estimates based on the matrix sampling design, two types of analysis problems were considered: estimating the population means of the split items; and regression analyses involving the split and core items. Properties of the no-imputation, multiple-imputation, and complete-data estimators across the 100 simulated data sets were compared with each other to assess bias and loss of efficiency due to matrix sampling combined with multiple imputation.

3.3.1 Estimating Population Means of Split Items

For the population mean of a split item, the simulated standardized bias of the no-imputation estimator was defined as $(\text{Ave}_{\text{NI}} - \text{Ave}_{\text{comp}}) / \text{SD}_{\text{NI}}$, where Ave_{NI} ,

Ave_{comp} , and SD_{NI} denote, respectively, the averages of the no-imputation and complete-data estimates and the standard deviation of no-imputation estimates across the 100 simulated data sets. An analogous simulated standardized bias was defined for the multiple-imputation (MI) estimator. Table 3 summarizes the simulated standardized biases for the 32 split items.

Table 3
Simulated Standardized Biases of the No-Imputation and Multiple-Imputation Estimators of the Population Means for the 32 Split Items

Standardized Biases	Frequency	
	No Imputation	Multiple Imputation
-1.4		1
(-1, -0.6]		4
(-0.6, -0.4]		5
(-0.4, -0.2]		4
(-0.2, 0)	15	10
(0, 0.2)	17	4
[0.2, 0.4]		
[0.4, 0.6]		2
[0.6, 1)		
1.4		1
4.6		1
Total	32	32

Because our matrix sampling mechanism results in missing data that are missing completely at random, the no-imputation estimators are close to unbiased. This is reflected in the simulation results by the fact that none of the absolute standardized biases is larger than 0.2. The multiple-imputation estimators generally have somewhat higher simulated standardized biases than do the no-imputation estimators, although the absolute standardized biases are less than one for all but three split items and less than 0.6 for all but seven. As a guideline for judging standardized biases, Cochran (1977, page 14) shows that a standardized bias of 0.6 produces nominal 95% confidence intervals having roughly 91% actual coverage. Any substantial biases observed in this study when matrix sampling is used in conjunction with multiple imputation are likely due to deficiencies in the imputation models and not to the matrix sampling itself, given that the no-imputation analyses were seen to be approximately unbiased. With larger sample sizes in an application to an actual survey, the corresponding standardized biases would tend to be moved upward because of the smaller denominators; but the standardized biases might also be moved downward because of improved large-sample approximations.

Loss of efficiency due to matrix sampling rather than using the full questionnaire can be assessed by comparing the sampling error of the no-imputation, multiple-imputation, and complete-data estimators (computed as standard deviations across the 100 simulated data sets). Table 4 summarizes the ratios of the simulated standard deviations

of the multiple-imputation estimators to those of the no-imputation estimators, and the ratios of the simulated standard deviations of the complete-data estimators to those of the multiple-imputation estimators (the term “simulated standard deviation” of an estimator is used rather than “simulated standard error” to avoid confusion with the estimated standard error that could be obtained from the analysis of each simulated data set).

Table 4

Ratios of the Simulated Standard Deviations of the No-Imputation (NI), Multiple-Imputation (MI), and Complete-Data (comp) Estimators of the Population Means for the 32 Split Items

Ratios	Frequency	
	SD _{MI} /SD _{NI}	SD _{comp} /SD _{MI}
(0.5, 0.6]		2
(0.6, 0.7]		9
(0.7, 0.8]		14
(0.8, 0.9]		6
(0.9, 0.95]	7	
(0.95, 1]	18	1
(1, 1.03]	7	
Total	32	32

Typically, the multiple-imputation estimators are more efficient than the no-imputation estimators, but the gain in efficiency is only modest, as indicated by the fact that most of the ratios SD_{MI}/SD_{NI} in Table 4 are between 0.9 and 1. Such modest gains in efficiency can be predicted roughly from the indices of predictive value based on data from NHANES II (displayed in Table 2), as follows. Because each split item is included in only half of the matrix sampling forms, it follows that the variance of a complete-data estimator of the mean of a split item should be about one-half the size of the variance of the corresponding no-imputation estimator. Dividing the numerator and denominator of expression (3) by V_{NI} , and setting $V_{comp}/V_{NI} = 0.5$, yields $2(1 - V_{MI}/V_{NI})$ as an approximate expression for the index of predictive value in this simulation study. For an index of 0.12, which is the median of the “Achieved Medium” indices in Table 2, it follows that V_{MI}/V_{NI} should be about 0.94. This ratio of variances is equivalent to a ratio of standard deviations of about $\sqrt{0.94} = 0.97$, which is near the middle of the range of ratios summarized in Table 4. In this study, because the multiple-imputation estimators are only modestly more efficient than the no-imputation estimators, and because the multiple-imputation estimators have some biases associated with them, the mean square errors for the multiple-imputation estimators are higher than those for the no-imputation estimators in 22 out of 32 cases.

The simulation results on the efficiency of the multiple-imputation estimators relative to the complete-data estimators also conform with theory. Since V_{comp}/V_{NI} should be about 0.5, and since V_{MI} should be slightly smaller than V_{NI} , it follows that V_{comp}/V_{MI} should be slightly larger

than 0.5, or equivalently, that the typical ratio of standard deviations SD_{comp}/SD_{MI} should be slightly larger than $\sqrt{0.5} = 0.71$. Indeed, the median of the ratios summarized in Table 4 is 0.75. An alternative to the multiple imputation estimation is two-phase weighting based on core item estimators and their differences between blocks. Any advantage in efficiency from multiple-imputation estimation would be due to the additional information from the split items.

3.3.2 Estimating Regression Coefficients

The matrix sampling and multiple-imputation methods were also evaluated for estimation of the coefficients of eight regression models, which were specified to be similar to models that have appeared in the literature. The regression models, which are listed in Table 5, had a total of 115 coefficients. No-imputation estimators for the regression coefficients were not included in the simulation study, although some theoretical results on their efficiency are discussed in this section.

For each regression coefficient, the simulated standardized bias was defined analogously to the definition used for each mean in Section 3.3.1. Table 6 summarizes the standardized biases for the 115 regression coefficients. Most of the standardized biases are small, with absolute values greater than one for only five coefficients and absolute values of 0.6 or greater for only seven.

Table 7 summarizes the ratios of the standard deviations of the complete-data estimates across the 100 simulated data sets to those of the multiple-imputation estimates, for the 115 regression coefficients. Separate summaries are displayed by whether the regression models involve split variables from only one block (Models 1, 2, 6, and 7) versus two blocks (Models 3, 4, 5, and 8). A larger proportion of the ratios are close to one than was the case for estimating means (Table 4). In addition, for several regression coefficients (particularly from Models 3, 6, 7, and 8), the simulated standard deviations of the complete-data estimators are moderately larger than those of the multiple-imputation estimators, and for one coefficient, the ratio is about two. Finally, there are four regression coefficients for which there appears to be a substantial loss of efficiency due to matrix sampling, with ratios less than 0.3 (one each from Models 1, 2, 5, and 8). The ratios close to or larger than one could be due in part to a lack of fit of some regression models to the complete data and a better fit of the models to the data completed by imputation, with the latter resulting from an imputation process that is based on regression models. Moreover, the two smallest ratios occur for regression models involving split variables from two blocks, for which the fraction of subjects in the matrix sample with no missing data is only one-sixth, as discussed further below.

Table 5
Regression Models Used in the Evaluation

Type of regression model	Dependent variable	Variables recoded to create predictors including interaction terms. Each model also includes an intercept term. For each variable, the number in the parentheses indicates the number of regression coefficients associated with the variable	Split variables in the regression models. For each variable, the number in the parentheses indicates the block containing the variable
1. Linear	G1P	HSSEX(1), HSAGEIR(1), DMARETHN(3), and GHP(1)	G1P(1)
2. Logistic	HAF10	HSSEX(1), HSAGEIR(1), DMARETHN(3), FEP(1), and BMPBMI(1)	FEP(3)
3. Logistic	HAF10	HSSEX(1), HSAGEIR(1), DMARETHN(3), HAD1(1), HAE3(1), PBP(1), FEP(1), CHP(1), and G1P(1)	FEP(3) and G1P(1)
4 and 5. Linear	SPPFVC	HSAGEIR(1), DMARETHN (3), HFA8R(2), and BMPBMI(1) [By gender (HSSEX), and restricted to never smokers (HAR1, HAR3)]	SPPFVC(1), HAR1(3), and HAR3(3)
6 and 7. Logistic	HCHP (1 IF CHP \geq 240 AND 0 OTHERWISE)	HSAGEIR(2), DMARETHN(3), HFA8R(1), BMPBMI(3), (HAR3,HAR1)(2), BMPBMI*HSAGEIR(6), and DMARETHN*BMPBMI(9) [By Gender (HSSEX)]	(HAR3, HAR1)(3)
8. Logistic	HAC1E	HSAGEIR(5), HSSEX(1), DMARETHN(3), BMPBMI(4), (HAR3, HAR1)(2), SPPPEAK(1), and SPPFVC(1)	HAC1E(2), (HAR1,HAR3)(3), SPPPEAK(3), and SPPFVC(2)

Table 6
Simulated Standardized Biases of the Multiple-Imputation Estimators for the 115 Regression Coefficients

Range of Standardized Biases	Frequency
-5.2	1
-1.5	1
-1.3	1
-1.1	1
(-1, -0.6]	2
(-0.6, -0.4]	2
(-0.4, -0.2]	3
(-0.2, 0)	52
(0, 0.2)	44
[0.2, 0.4)	6
[0.4, 0.6)	1
[0.6, 1)	
3.7	1
Total	115

Table 7
Ratios of the Simulated Standard Deviations of the Complete-Data Estimators to those of the Corresponding Multiple-Imputation Estimators, for the 115 Regression Coefficients, by Whether the Regression Models Involve Split Variables from Only One Block Versus Two Blocks

Range of Ratios	Frequency	
	One Block	Two Blocks
(0, 0.1]		1
(0.1, 0.2]		1
(0.2, 0.3]	2	
(0.3, 0.4]		
(0.4, 0.5]	1	
(0.5, 0.6]		3
(0.6, 0.7]	2	3
(0.7, 0.8]	2	7
(0.8, 0.9]	4	4
(0.9, 0.95]	2	3
(0.95, 1]	29	8
(1, 1.05]	20	5
(1.05, 1.1]	4	2
(1.1, 1.2]	3	2
(1.2, 1.4]		4
(1.4, 1.6]		2
2.0		1
Total	69	46

For regression models involving split variables from only one block, the theoretical efficiency of the complete-data estimator relative to the no-imputation estimator, that is, the ratio of the variance of the latter to the former, is approximately two because only half of the subjects in the matrix sample will have complete data on those variables; and for regression models involving split variables from two blocks, the theoretical relative efficiency is approximately six. In contrast, the respective simulated relative efficiencies of the complete-data estimator relative to the multiple-imputation estimator, that is, the inverses of the squared ratios summarized in Table 7, are less than two for 64 out of 69 coefficients when only one block is involved; and they are less than six for 44 out of 46 coefficients when two blocks are involved. Thus, the multiple-imputation estimators are generally more efficient than the no-imputation estimators for regression problems. Nevertheless, the large losses of efficiency of the multiple-imputation estimators relative to the complete-data estimators for some coefficients as well as the apparent gains in efficiency for other coefficients are worth further investigation.

3.4 Additional Limitations of the Simulation Study

This section briefly discusses some additional limitations of the simulation study and adjustments required during the implementation of the study.

Originally, two questions about two conditions, gout and lupus (HAC1M and HAC1L) were designated as split items. Due to low prevalence of these two conditions in the constructed finite population, many of the simulated samples had no subjects with these conditions. After a few preliminary runs, the designations for these two items were changed from split to core. In general, in situations with limited sample sizes, conditions with very low prevalence rates may need to be designated as core items. In addition, due to issues such as some split items appearing in NHANES III but not in NHANES II, as well as logical linkages between some split items, the number of split items per block in the simulation study varied slightly more than intended (from 6 to 10).

In the regression models listed in Table 3, the number of predictors ranged from 8 to 27, because some of the regression models included interaction terms as predictors. Even with the sample size of 1,200 for the simulated samples, some of the complete-data estimators were unstable. This was due in part to small sample sizes for some combinations of variables that affected the estimation of interactions. Note that in many applications of matrix sampling to large surveys, complete-data sample sizes would be substantially larger than the size of 1,200 used in our simulation study.

The Monte Carlo standard errors of the simulated averages in this study are approximately one-tenth of the standard deviations of the individual quantities across the 100 samples. However, the standard deviations across the

samples varied widely from one estimand to another, due to differences in scaling. For example, the simulated standard deviations of the complete-data estimators of the 115 regression coefficients ranged from 9.8×10^{-5} to 1169.6. More precise estimates of bias and efficiency could be computed based on a larger number of simulated samples than was used in this study.

4. Discussion

In this paper, a method was developed for creating matrix sampling designs that have the property that the items included on forms are predictive of the items that have been excluded. The feasibility of implementing such designs in a complex, large-scale health survey was demonstrated via an example involving the National Health and Nutrition Examination Survey. Matrix sampling designs, in conjunction with multiple imputation, can be used to expand the scope of a survey without increasing respondent burden or unduly increasing the burden to subsequent data analysts.

In the study involving NHANES data, the multiple-imputation analyses of data from the matrix samples were modestly effective, with minor evidence of bias and with greater efficiency than simply analyzing the matrix sampled data without imputation. The increased efficiency was especially evident in the context of regression analyses.

Matrix sampling in the NHANES example typically resulted in large losses of precision compared to what could have been achieved with a longer, complete survey (*i.e.*, no matrix sampling), however. This finding, which is in contrast with the more promising results obtained in other applications of matrix sampling, highlights the importance of including good predictors of the split items in a survey. For example, an application of matrix sampling to an educational survey (*e.g.*, Beaton and Zwick 1992) has been much more successful, because the split items are highly correlated responses to questions designed to measure the same trait. Raghunathan and Grizzle (1995) also demonstrated much greater recovery of information on omitted items in the context of a health survey.

The items in the NHANES example were chosen mainly to represent a variety of important health characteristics, without much consideration given to their ability to predict or be predicted by other variables. Many of the split items represented rare illnesses that are not well predicted by common medical conditions and standard laboratory measurements; in hindsight, these variables were not good candidates for split items. Variables representing rare events may also cause difficulties with many common statistical methods that rely on large-sample approximations, making them less amenable to model-based imputation.

Better candidates for matrix sampling designs are "panels" of inter-related items. For example, matrix sampling techniques can be useful when there are multiple measurements of the same (or closely related) quantities, and it is desired to collect some of the measurements for subsets of the survey respondents due to cost and time considerations. Some rudimentary forms of matrix sampling are already being applied in such settings, and there may be substantial improvements possible by applying methods, such as those developed in this paper, that aim to exploit the associations among the variables.

Acknowledgments

The work of Neal Thomas and Trivellore E. Raghunathan was supported in part by a professional services contract between NCHS and Datametrics Research, Inc. The authors thank Randy Curtin of NCHS for helpful suggestions. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

References

- Beaton, A., and Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics*, 17, 95-109.
- Cochran, W.G. (1977). *Sampling Techniques*. Third Edition. New York: John Wiley & Sons, Inc.
- Houseman, E., and Milton, D. (2006). Partial questionnaire designs, questionnaire nonresponse, and attributable fraction: Applications to adult onset asthma. *Statistics in Medicine*, 25, 1499-1519.
- Kennickell, A.B. (1991). Imputation of the 1989 Survey of Consumer Finances: stochastic relaxation and multiple imputation. *Proceedings the Survey Research Methods Section*, American Statistical Association, 112-121.
- McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models*. Second Edition, London: Chapman Hall.
- Meng, X.-L. (1994). Multiple imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 9, 538-573.
- Navarro, A., and Griffin, R. (1993). Matrix sampling designs for the year 2000 Census. *Proceedings the Survey Research Methods Section*, American Statistical Association, 480-485.
- Oudshoorn, K., Van Buuren, S. and Van Rijckevorsel, J. (1999). Flexible multiple imputation by chained equations of the AVO-95 Survey. Leiden: TNO Prevention and Health, Report PG/VGZ/99.045.
- Raghunathan, T.E., and Grizzle, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54-63.

- Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85-95.
- Rubin, D.B. (1976). Inference and missing data (with discussion). *Biometrika*, 63, 581-592.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.
- Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.
- Rubin, D.B., and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- Schafer, J.L., and Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association*, 95, 144-154.
- Schenker, N., Gentleman, J.F., Rose, D., Hing, E. and Shimizu, I.M. (2002). Combining estimates from complementary surveys: A case study using prevalence estimates from national health surveys of households and nursing homes. *Public Health Reports*, 117, 393-407.
- Shoemaker, D.M. (1973). *Principles and Procedures of Matrix Sampling*. Cambridge, MA: Ballinger.
- Sirotnik, K., and Wellington, R. (1977). Incidence sampling: An integrated theory for matrix sampling. *Journal of Educational Measurement*, 14, 343-399.
- Wacholder, S., Carroll, R.J., Pee, D. and Gail, M.H. (1994). The partial questionnaire design for case-control studies (with discussion). *Statistics in Medicine*, 13, 623-649.
- Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- Zeger, L.M., and Thomas, N. (1997). Efficient matrix sampling for correlated latent traits: Examples from the National Assessment of Educational Progress. *Journal of the American Statistical Association*, 92, 416-425.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2006.

- J.-F. Beaumont, *Statistics Canada*
 Y. Berger, *The University of Reading, UK*
 C. Boudreau, *Medical College of Wisconsin*
 L. Burck, *Central Bureau of Statistics, Israël*
 F. Butar, *Sam Houston University*
 J. Chipperfield, *Australian Bureau of Statistics*
 S.R. Chowdhury, *Westat Inc.*
 G. Datta, *University of Georgia*
 P. Duchesne, *Université de Montréal*
 K. Duncan, *Dominican University, Chicago*
 F. Dupont, *INSEE*
 G.B. Durrant, *Southampton Statistical Sciences Research Institute, University of Southampton, UK*
 M. Elliott, *University of Michigan*
 J. Eltinge, *United States Bureau of Labor Statistics*
 M. Feder, *Research Triangle Institute*
 R. Folsom, *Research Triangle Institute*
 O. Frank, *Stockholm University*
 J. Gambino, *Statistics Canada*
 C. Girard, *Statistics Canada*
 M. Gosh, *University of Florida*
 B. Graubard, *National Cancer Institute*
 G. Griffiths, *Australian Bureau of Statistics*
 D. Haziza, *Statistics Canada*
 J. Horgan, *Dublin City University*
 V.G. Iannacchione, *RTI International*
 J. Jiang, *University of California at Davis*
 J.-K. Kim, *Department of Applied Statistics, Korea*
 P. Kokic, *Australian Bureau of Agriculture and Resource Economics*
 P. Kott, *United States Department of Agriculture*
 M. Kovačević, *Statistics Canada*
 F. Kreuter, *Joint Program in Survey Methodology*
 M.D. Larsen, *Iowa State University*
 P. Lavallée, *Statistics Canada*
 H. Lee, *Westat, Inc.*
 R. Lehtonen, *University of Helsinki*
 C. Leon, *Statistics Canada*
 W.W. Lu, *Department of Mathematics and Statistics*
 A. Matei, *Université de Neuchâtel, Suisse*
 D. Melec, *United States Bureau of the Census*
 J.M. Montaquila, *Westat, Inc.*
 R. Munnich, *University of Tubingen*
 J. Opsomer, *Iowa State University*
 Z. Patak, *Statistics Canada*
 D. Pfeiffermann, *Israël and University of Southampton*
 N. Prasad, *University of Alberta*
 L. Qualité, *Université de Neuchâtel, Suisse*
 M.G. Ranalli, *Universita' degli Studi di Perugia*
 J.N.K. Rao, *Carleton University*
 L.-P. Rivest, *Université Laval*
 O. Sautory, *Insee-Cepe*
 J. Schafer, *Pennsylvania State University*
 A. Scott, *University of Auckland*
 R. Singh, *U.S. Census Bureau*
 C. Skinner, *University of Southampton*
 E. Stuart, *Mathematica Policy Research Inc*
 C.J. Swartz, *Simon Fraser University*
 R. Valliant, *University of Michigan*
 Z. Wang, *Wilfrid Laurier University, Waterloo*
 M. Winglee, *Westat*
 C. Wu, *University of Waterloo*
 W. Yung, *Statistics Canada*
 E. Zanutto, *Department of Statistics, The Wharton School, University of Pennsylvania*

Acknowledgements are also due to those who assisted during the production of the 2006 issues: Cécile Bourque, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Micheal Pelchat and Isabelle Poliquin (Dissemination Division), Nadine Lacroix (Client Services Division), Sheri Buck (Systems Development Division), François Beaudin (Official Languages and Translation Division) and Sophie Chartier (Business Survey Methods Division). Finally we wish to acknowledge Christine Cousineau, Céline Ethier and Denis Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 22, No. 1, 2006

Frequency Domain Analyses of SEATS and X-11/12-ARIMA Seasonal Adjustment Filters for Short and Moderate-Length Time Series David F. Findley and Donald E.K. Martin	1
Variance Estimation by Jackknife Method Under Two-Phase Complex Survey Design Debesh Roy and Md. Safiquzzaman.....	35
Estimating the Undercoverage of a Sampling Frame Due to Reporting Delays Dan Hedlin, Trevor Fenton, John W. McDonald, Mark Pont, and Suojin Wang.....	53
Raking Ratio Estimation: An Application to the Canadian Retail Trade Survey Michael A. Hidirolou and Zdenek Patak	71
Survey Estimation Under Informative Nonresponse with Follow-up Seppo Laaksonen and Ray Chambers.....	81
An Analysis of the Relationship Between Survey Burden and Nonresponse: If We Bother Them More, Are They Less Cooperative? Jaki Stanley McCarthy, Daniel G. Beckler, and Suzette M. Qualey	97
How the United States Measures Well-being in Household Surveys Daniel H. Weinberg.....	113
Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints Marcello D'Orazio, Marco Di Zio, and Mauro Scanu.....	137
Erratum	159
Book and Software Reviews	161
In Other Journals	173

Volume 22, No. 2, 2006

Preface.....	iii
Putting a Questionnaire on the Web is not Enough - A Comparison of Online and offline Surveys Conducted in the Context of the German Federal Election 2002 Thorsten Faas and Harald Schoen.....	177
An Experimental Study on the Effects of Personalization, Survey Length Statements, Progress Indicators, and Survey Sponsor Logos in Web Surveys Dirk Heerwegh and Geert Loosveldt	191
Dual Frame Web - Telephone Sampling for Rare Groups Edward Blair and Johnny Blair	211
Merely Incidental?: Effects of Response Format on Self-reported Behavior Randall K. Thomas and Jonathan D. Klein	221
Use and Non-use of Clarification Features in Web Surveys Frederick G. Conrad, Mick P. Couper, Roger Tourangeau, and Andrey Peytchev.....	245
The Influence of Web-based Questionnaire Presentation Variations on Survey Cooperation and Perceptions of Survey Quality Jill T. Walston, Robert W. Lissitz, and Lawrence M. Rudner	271
Can Web and Mail Survey Modes Improve Participation in an RDD-based National Health Surveillance? Michael W. Link and Ali Mokdad.....	293
Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey Mirta Galesic.....	313
Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys Sunghee Lee.....	329
Book and Software Review	351

All inquiries about submissions and subscriptions should be directed to jos@scb.se

Volume 34, No. 2, June/juin 2006

Louis-Paul RIVEST & Ted CHANG Regression and correlation for 3×3 rotation matrices	187
Christian BOUDREAU & Jerald F. LAWLESS Survival analysis based on the proportional hazards model and survey data	203
Edit GOMBAY & Abdulkadir HUSSEIN A class of sequential tests for two-sample composite hypotheses	217
Donald L. MCLEISH & Cynthia A. STRUTHERS Estimation of regression parameters in missing data problems	233
Sanjoy K. SINHA Robust inference in generalized linear models for longitudinal data	261
Xiaogang WANG Approximating Bayesian inference by weighted likelihood	279
Borek PUZA & Terence O'NEILL Interval estimation via tail functions	299
M. Farid ROHANI, Khalil SHAFIE & Siamak NOORBALOOCHI A Bayesian signal detection procedure for scale-space random fields	311
Marlos A.G. VIANA & Hak-Myung LEE Correlation analysis of ordered symmetrically dependent observations and their concomitants of order statistics	327
Kanchan MUKHERJEE Pseudo-likelihood estimation in ARCH models	341
Forthcoming papers/Articles à paraître	357
Online access to The Canadian Journal of Statistics	358
Services en ligne de La revue canadienne de statistique	358

Volume 34, No. 3, September/septembre 2006

Changbao WU & J.N.K. RAO Pseudo-empirical likelihood ratio confidence intervals for complex surveys	359
Paul GUSTAFSON, Shahadut HOSSAIN & Ying C. MACNAB Conservative prior distributions for variance parameters in hierarchical models	377
Jinhong YOU, Yong ZHOU & Gemai CHEN Corrected local polynomial estimation in varying-coefficient models with measurement errors	391
José T.A.S. FERREIRA & Mark F.J. STEEL On describing multivariate skewed distributions: a directional approach	411
Fabienne COMTE, Yves ROZENHOLC & Marie-Luce TAUPIN Penalized contrast estimator for adaptive density deconvolution	431
Jonathan B. HILL Strong orthogonal decompositions and non-linear impulse response functions for infinite-variance processes	453
Jean-Michel LOUBES, Élie MAZA, Marc LAVIELLE & Luis RODRÍGUEZ Road trafficking description and short term travel time forecasting, with a classification method	475
Sylvia R. ESTERBY Variables related to codling moth abundance and the efficacy of the Okanagan Sterile Insect Release Program	493
Bob VERNON, Howard THISTLEWOOD, Scott SMITH & Todd KABALUK A GIS application to improve codling moth management in the Okanagan Valley of British Columbia	494
Farouk NATHOO, Laurie AINSWORTH, Paramjit GILL & Charmaine B. DEAN Codling moth incidence in Okanagan orchards	500
Gaétan DAIGLE, Thierry DUCHESNE, Emmanuelle RENY-NOLIN & Louis-Paul RIVEST Étude de l'influence de la topographie et des caractéristiques des vergers sur l'efficacité du programme d'épandage d'insectes stériles pour le carpocapse de la pomme (<i>Laspeyresia pomonella</i>)	511
Sylvia R. ESTERBY, Howard THISTLEWOOD, Bob VERNON & Scott SMITH Analysis of codling moth data from the Okanagan Sterile Insect Release Program	521
Forthcoming papers / Articles à paraître	531
Volume 35 (2007): Subscription rates / Frais d'abonnement	532
Online access to The Canadian Journal of Statistics	533

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, N° 1) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version papier pourrait être requise pour les formules et graphiques.

1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8 1/2 par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp()$ et $\log()$ etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω , \circ , O, 0; l, 1).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.
4. **Figures et tableaux**
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. **Bibliographie**
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).
 - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.
6. **Communications brèves**

Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.

Changhao WU & J.N.K. RAO	Pseudo-empirical likelihood ratio confidence intervals for complex surveys	359
Paul GUSTAFSSON, Shahadut HOSSAIN & Ying C. MACNAB	Conservative prior distributions for variance parameters in hierarchical models	377
Jinlong YOU, Yong ZHOU & Gemai CHEN	Corrected local polynomial estimation in varying-coefficient models with measurement errors	391
José T.A.S. FERREIRA & Mark F.J. STEEL	On describing multivariate skewed distributions: a directional approach	411
Fabienne COMTE, Yves ROZENHOLC & Martine-Luce TAPPIN	Penalized contrast estimator for adaptive density deconvolution	431
Jonathan B. HILL	Strong orthogonal decompositions and non-linear impulse response functions for infinite-variance processes	453
Jean-Michel LOUBES, Elie MAZA, Marc LAVIELLE & Luis RODRIGUEZ	Road trafficking description and short term travel time forecasting, with a classification method	475
Sylvia R. ESTERBY	Variables related to codling moth abundance and the efficacy of the Okanagan Sterile Insect Release Program	493
Bob VERNON, Howard THISTLEWOOD, Scott SMITH & Todd KABALUK	A GIS application to improve codling moth management in the Okanagan Valley of British Columbia	494
Farouk NATHOO, Laurie AINSWORTH, Paramjit GILL & Charmaine B. DEAN	Codling moth incidence in Okanagan orchards	500
Gaëtan DAIGLE, Thierry DUCHESSNE, Emmanuelle RENY-NOLIN & Louis-Paul RIVEST	Etude de l'influence de la topographie et des caractéristiques des vergers sur l'efficacité du programme d'épandage d'insectes stériles pour le carpocapse de la pomme (<i>Laspeyresia pomonella</i>)	511
Sylvia R. ESTERBY, Howard THISTLEWOOD, Bob VERNON & Scott SMITH	Analysis of codling moth data from the Okanagan Sterile Insect Release Program	521
	Forthcoming papers / Articles à paraître	531
	Volume 35 (2007): Subscription rates / Frais d'abonnement	532
	Online access to The Canadian Journal of Statistics	533

Volume 34, No. 2, June/juin 2006

Louis-Paul RIVEST & Ted CHANG	187
Regression and correlation for 3×3 rotation matrices	
Christian BODRÉAU & Jerald F. LAWLESS	203
Survival analysis based on the proportional hazards model and survey data	
Edit GOMBAY & Abdulkadir HUSSEIN	217
A class of sequential tests for two-sample composite hypotheses	
Donald L. MCLEISH & Cynthia A. STRUTHERS	233
Estimation of regression parameters in missing data problems	
Sanjoy K. SINHA	261
Robust inference in generalized linear models for longitudinal data	
Xiaogang WANG	279
Approximating Bayesian inference by weighted likelihood	
Borek PUZA & Terence ONEILL	299
Interval estimation via tail functions	
M. Fatih ROHANI, Khalil SHAFIE & Siamak NOORBALOOCHI	311
A Bayesian signal detection procedure for scale-space random fields	
Marios A.G. VIANÀ & Hak-Myun LEE	327
Correlation analysis of ordered symmetrically dependent observations and their concomitants of order statistics	
Kanchan MUKHERJEE	341
Pseudo-likelihood estimation in ARCH models	
Forthcoming papers/Articles à paraître	357
Online access to The Canadian Journal of Statistics	358
Services en ligne de La revue canadienne de statistique	358

Preface.....	iii
Putting a Questionnaire on the Web is not Enough - A Comparison of Online and offline Surveys Conducted in the Context of the German Federal Election 2002	177
Thorsten Faas and Harald Schoen.....	177
An Experimental Study on the Effects of Personalization, Survey Length Statements, Progress Indicators, and Survey Sponsor Logos in Web Surveys	191
Dirk Heerwegh and Geert Loosveldt.....	191
Dual Frame Web - Telephone Sampling for Rare Groups	211
Edward Blair and Johnny Blair.....	211
Merely Incidental?: Effects of Response Format on Self-reported Behavior	221
Randall K. Thomas and Jonathan D. Klein.....	221
Use and Non-use of Clarification Features in Web Surveys	245
Frederick G. Conrad, Mick P. Couper, Roger Tourangeau, and Andrey Peytchev.....	245
The Influence of Web-based Questionnaire Presentation Variations on Survey Cooperation and Perceptions of Survey Quality	271
Jill T. Walston, Robert W. Lissitz, and Lawrence M. Rudner.....	271
Can Web and Mail Survey Modes Improve Participation in an RDD-based National Health Surveillance?	293
Michael W. Link and Ali Mokdad.....	293
Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey	313
Mirta Galeisic.....	313
Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys	329
Sunghye Lee.....	329
Book and Software Review.....	351

All inquiries about submissions and subscriptions should be directed to jos@scb.se

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

JOURNAL OF OFFICIAL STATISTICS
An International Review Published by Statistics Sweden

Contents
Volume 22, No. 1, 2006

Frequency Domain Analyses of SEATS and X-11/12-ARIMA Seasonal Adjustment Filters for Short and Moderate-Length Time Series	David F. Findley and Donald E.K. Martin	1
Variance Estimation by Jackknife Method Under Two-Phase Complex Survey Design	Debash Roy and Md. Saifuzzaman	35
Estimating the Undercoverage of a Sampling Frame Due to Reporting Delays	Dan Hedlin, Trevor Fenton, John W. McDonald, Mark Pont, and Suolin Wang	53
Raking Ratio Estimation: An Application to the Canadian Retail Trade Survey	Michael A. Hidiroglou and Zdenek Patak	71
Survey Estimation Under Informative Nonresponse with Follow-up	Seppo Laaksonen and Ray Chambers	81
An Analysis of the Relationship Between Survey Burden and Nonresponse: If We Bother Them More, Are They Less Cooperative?	Jaki Stanley McCarthy, Daniel G. Becker, and Suzette M. Qualey	97
How the United States Measures Well-being in Household Surveys	Daniel H. Weinberg	113
Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints	Marcello D'Orazio, Marco Di Zio, and Mauro Scannu	137
Erratum		159
Book and Software Reviews		161
In Other Journals		173

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2006.

J.-F. Beaumont, *Statistique Canada*
 J.-F. Berger, *The University of Reading, UK*
 C. Boudreau, *Medical College of Wisconsin*
 L. Burck, *Central Bureau of Statistics, Israel*
 F. Butar, *Sam Houston University*
 I. Chipperfield, *Australian Bureau of Statistics*
 S.R. Chowdhury, *Westat Inc.*
 G. Data, *University of Georgia*
 P. Duchesne, *Université de Montréal*
 K. Duncan, *Dominican University, Chicago*
 F. Dupont, *INSEE*
 G.B. Durrant, *Southampton Statistical Sciences Research Institute, University of Southampton, UK*
 M. Elliott, *University of Michigan*
 J. Elling, *United States Bureau of Labor Statistics*
 M. Feder, *Research Triangle Institute*
 R. Folsom, *Research Triangle Institute*
 O. Frank, *Stockholm University*
 J. Gambino, *Statistique Canada*
 C. Girard, *Statistique Canada*
 M. Gosh, *University of Florida*
 B. Graubard, *National Cancer Institute*
 D. Griffiths, *Australian Bureau of Statistics*
 G. Haziza, *Statistique Canada*
 J. Horgan, *Dublin City University*
 V.G. Iannacchione, *RTI International*
 J. Jiang, *University of California at Davis*
 J.-K. Kim, *Department of Applied Statistics, Korea*
 P. Kokic, *Australian Bureau of Agriculture and Resource Economics*
 P. Kott, *United States Department of Agriculture*
 M. Kovacevich, *Statistique Canada*
 F. Kreuter, *Joint Program in Survey Methodology*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2006: Cécile Bourque, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanotte, Michael Pelchat et Isabelle Poliquin (Division de la diffusion), Nadine Lacroix (Division des services à la clientèle), Sheri Buck (Division du développement de systèmes), François Beaudin (Division des langues officielles et traduction) et Sophie Charrier (Division des méthodes d'enquêtes auprès des entreprises). Finalement nous désirons exprimer notre reconnaissance à Christine Cousineau, Céline Ethier et Denis Lemire de la Division des méthodes d'enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

M.D. Larsen, *Iowa State University*
 P. Lavallee, *Statistique Canada*
 H. Lee, *Westat, Inc.*
 R. Lehtonen, *University of Helsinki*
 C. Leon, *Statistique Canada*
 W.W. Lu, *Department of Mathematics and Statistics*
 A. Matei, *Université de Neuchâtel, Suisse*
 D. Melec, *United States Bureau of the Census*
 J.M. Montaquila, *Westat, Inc.*
 R. Munich, *University of Tübingen*
 J. Opsomer, *Iowa State University*
 Z. Patak, *Statistique Canada*
 D. Pfeffermann, *Israel and University of Southampton*
 N. Prasad, *University of Alberta*
 L. Qualité, *Université de Neuchâtel, Suisse*
 M.G. Ranalli, *Università degli Studi di Perugia*
 J.N.K. Rao, *Carleton University*
 L.-P. Rivest, *Université Laval*
 O. Sautory, *Insee-Cepe*
 J. Schater, *Pennsylvania State University*
 A. Scott, *University of Auckland*
 R. Singh, *U.S. Census Bureau*
 C. Skinner, *University of Southampton*
 E. Stuart, *Mathematica Policy Research Inc*
 C.J. Swartz, *Simon Fraser University*
 R. Valliam, *University of Michigan*
 Z. Wang, *Wilfrid Laurier University, Waterloo*
 M. Wingler, *Westat*
 C. Wu, *University of Waterloo*
 W. Yung, *Statistique Canada*
 E. Zanutto, *Department of Statistics, The Wharton School, University of Pennsylvania*

- Raghunathan, T.E., et Gutzle, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54-63.
- Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. et Solenberger, P. (2001). Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Techniques d'enquête*, 27, 91-103.
- Rubin, D.B. (1976). Inference and missing data (avec discussion). *Biometrika*, 63, 581-592.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York : John Wiley & Sons, Inc.
- Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.
- Rubin, D.B., et Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- Schaffer, J.L., et Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association*, 95, 144-154.
- Schenker, N., Gentleman, J.F., Rose, D., Hing, E. et Shimizu, L.M. (2002). Combining prevalence estimates from national health surveys case study using prevalence estimates from complementary surveys of households and nursing homes. *Public Health Reports*, 117, 393-407.
- Shoenaker, D.M. (1973). *Principles and Procedures of Matrix Sampling*. Cambridge, MA : Ballinger.
- Stromik, K., et Wellington, R. (1977). Incidence sampling: An integrated theory for matrix sampling. *Journal of Educational Measurement*, 14, 343-399.
- Wacholder, S., Carroll, R.J., Pez, D. et Gail, M.H. (1994). The partial questionnaire design for case-control studies (avec discussion). *Statistics in Medicine*, 13, 623-649.
- Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York : Springer-Verlag.
- Zeger, L.M., et Thomas, N. (1997). Efficient matrix sampling for correlated latent traits: Examples from the National Assessment of Educational Progress. *Journal of the American Statistical Association*, 92, 416-425.

propriété de produire des questionnaires contenant des questions qui sont prédictives de celles qui ont été exclues. La faisabilité de l'application de plans de ce genre à une enquête sur la santé complexe, à grande échelle, est démontrée au moyen d'un exemple portant sur la National Health and Nutrition Examination Survey. Il est possible d'utiliser des plans d'échantillonnage matriciel, conjugués à l'imputation multiple, pour étendre la portée d'une enquête, sans accroître le fardeau de réponse, ni sans augmenter excessivement le fardeau subéquent des analystes des données.

Dans l'étude portant sur les données de la NHANES, les analyses en présence d'imputation multiple des données provenant des échantillons matriciels étaient modérément efficaces, les preuves de biais étant minimes et l'efficacité, plus grande que celle de l'analyse des données provenant des échantillons matriciels uniquement, sans imputation. Le gain d'efficacité est particulièrement évident dans le contexte des analyses de régression.

Cependant, dans l'exemple fondé sur la NHANES, l'échantillonnage matriciel entraînait généralement une grande perte de précision comparativement aux résultats que l'on aurait obtenus au moyen d'un questionnaire complet, plus long (c'est à dire sans échantillonnage matriciel). Cette constatation, qui est en contradiction avec les résultats plus prometteurs obtenus dans les autres applications de l'échantillonnage matriciel, fait ressortir combien il est important d'intégrer de bons prédicteurs des questions échantillonnées dans un questionnaire. Par exemple, une application de l'échantillonnage matriciel à une enquête sur les acquis scolaires (par exemple, Beaton et Zwick 1992) a donné de

échantillonnées étaient fortement corrélées à des questions nettement meilleurs résultats, parce que les questions conçues pour mesurer le même trait. Raghunathan et Grizzle (1995) ont également donné la preuve d'un recouvrement beaucoup plus important de l'information sur les questions omises dans le contexte d'une enquête sur la santé.

Dans l'exemple de la NHANES, les questions ont été choisies principalement en vue de représenter une gamme de caractéristiques importantes de la santé, sans trop tenir compte de leur capacité à prédire ou à être prédites par d'autres variables. Un grand nombre de questions échantillonnées correspondaient à des maladies rares qui ne sont pas bien prédites par les problèmes de santé courants et les mesures standard de laboratoire; rétrospectivement, ces variables n'étaient pas de bonnes candidates pour les questions échantillonnées. Les variables représentant des événements rares peuvent aussi poser des difficultés dans le cas de nombreuses méthodes statistiques courantes qui s'appuient sur des approximations en grand échantillon, de sorte qu'elles se prêtent moins bien à l'imputation fondée sur un modèle.

Remerciements

Les « panels » de questions interdépendantes sont de meilleurs candidats à l'échantillonnage matriciel. Par exemple, les techniques d'échantillonnage matriciel peuvent être utilisées dans des situations où sont faites des mesures multiples des mêmes quantités (ou de quantités étroitement liées), et où il est souhaitable de recueillir certaines mesures auprès de sous-ensembles des répondants à l'enquête à cause de contraintes de coûts et de temps. Certains formes rudimentaires d'échantillonnage matriciel sont déjà appliquées dans de telles circonstances et des améliorations sensibiles pourraient être réalisées en appliquant des méthodes, telles que celle élaborée dans le présent article, qui visent à exploiter les associations entre les variables.

Bibliographie

Les travaux de Neal Thomas et de Trivellore E. Raghunathan ont été financés en partie par un contrat de services professionnels conclu par le NCHS et Diarmetrics Research, Inc. Les auteurs remercient Randy Curtin du NCHS pour ses conseils utiles. Les résultats et les conclusions présentés dans le présent article sont ceux des auteurs et ne représentent pas forcément les opinions du National Center for Health Statistics des Centers for Disease Control and Prevention.

Beaton, A., et Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics*, 17, 95-109.

Cochran, W.G. (1977). *Sampling Techniques*. Troisième édition. New York : John Wiley & Sons, Inc.

Housemann, E., et Million, D. (2006). Partial questionnaire designs, questionnaire nonresponse, and attributable fraction: Applications to adult onset asthma. *Statistics in Medicine*, 25, 1499-1519.

Kennickell, A.B. (1991). Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. *Proceedings the Survey Research Methods Section, American Statistical Association*, 112-121.

McCullagh, P., et Nelder, J. (1989). *Generalized Linear Models*. Deuxième édition, London : Chapman Hall.

Meng, X.-L. (1994). Multiple imputation inferences with uncongenial sources of input (avec discussion). *Statistical Science*, 9, 538-573.

Navarro, A., et Griffin, R. (1993). Matrix sampling designs for the year 2000 Census. *Proceedings the Survey Research Methods Section, American Statistical Association*, 480-485.

Oudshoorn, K., Van Buuren, S. et Van Rijkevorsel, J. (1999). Flexible multiple imputation by chained equations of the AV-O-95 Survey. Leiden: TNO Prevention and Health, Rapport PG/VGZ/99.045.

Tableau 7
Ratios des écarts-types simulés des estimateurs sur les données complètes à ceux des estimateurs en présence d'imputation correspondants, pour les 15 coefficients de régression, selon que les modèles de régression comportent des variables échantillonnées provenant d'un seul bloc ou de deux blocs

Fréquence		Fourchette des ratios	
Un bloc	Deux blocs	Un bloc	Deux blocs
1	1	(0, 0,1]	(0, 0,1]
2	2	(0,1, 0,2]	(0,2, 0,3]
1	1	(0,3, 0,4]	(0,4, 0,5]
3	3	(0,5, 0,6]	(0,6, 0,7]
2	2	(0,7, 0,8]	(0,8, 0,9]
4	4	(0,9, 0,95]	(0,95, 1]
2	2	(1, 1,05]	(1,05, 1,1]
3	3	(1,1, 1,2]	(1,2, 1,4]
2	2	(1,4, 1,6]	2,0
46	69	Total	

3.4 Limites supplémentaires de l'étude par simulation

À la présente section, nous discutons brièvement de certaines limites supplémentaires de l'étude par simulation et des ajustements qui ont été requis durant sa mise en œuvre.

Au départ, deux questions au sujet de deux problèmes de santé, la goutte et le lupus (HACIM et HACIL) ont été désignées comme questions échantillonnées. À cause de la faible prévalence de ces deux problèmes de santé dans la population finie construite, un grand nombre des échantillons simulés ne contenaient pas de sujets qui en étaient atteints. Après quelques passages préliminaires, nous avons modifié la désignation de ces deux questions et en avons fait des questions communes. En général, dans les situations où les tailles d'échantillon sont limitées, il pourrait être nécessaire de considérer les questions sur les problèmes de santé dont la prévalence est très faible comme des questions communes. En outre, à cause de problèmes tels que l'existence de questions échantillonnées dans la NHANES III, mais pas dans la NHANES II, ainsi que l'existence de liens logiques entre certaines questions échantillonnées, le nombre de ces dernières par bloc dans l'étude de simulation variait un peu plus que prévu (de 6 à 10).

Dans les modèles de régression énumérés au tableau 3, le nombre de prédicteurs variaient de 8 à 27, parce que certains modèles comprenaient des termes d'interaction en tant que

prédicteur. Même avec une taille de 1 200 pour les échantillons simulés, certains estimateurs sur données complètes étaient instables. Cela était dû en partie à la petite taille d'échantillon pour certaines combinaisons de variables qui affectaient l'estimation des interactions. Il convient de souligner que, dans de nombreuses applications de l'échantillonnage matriciel à de grandes enquêtes, les tailles d'échantillon avec données complètes seraient sensiblement plus grandes que celle de 1 200 utilisée dans notre étude par simulation.

Les erreurs-types de Monte Carlo des moyennes simulées dans la présente étude sont égales au dixième environ des écarts-types des quantités individuelles sur l'ensemble des 100 échantillons. Cependant, les écarts-types sur l'ensemble des échantillons variaient fortement d'un paramètre estimé à l'autre, à cause de différences d'échelonnage. Par exemple, les écarts-types simulés des estimateurs sur données complètes des 115 coefficients de régression variaient de $9,8 \times 10^{-5}$ à 1 169,6. Des estimations plus précises du biais et de l'efficacité pourraient être calculées en se fondant sur un plus grand nombre d'échantillons simulés que celui utilisé dans l'étude.

4. Discussion

Le présent article décrit l'élaboration d'une méthode en vue de créer des plans d'échantillonnage matriciel ayant la

Tableau 5
Modèles de régression utilisés dans l'évaluation

Type de modèle de régression	Variable dépendante	Variables recodées pour créer des variables échantillonées dans les modèles de régression. Pour chaque variable, le chiffre entre parenthèses indique le bloc contenant la variable
1. Linéaire	G1P	HSSEX(1), HSAGEIR(1), DMARETHN(3), et GHP(1)
2. Logistique	HAF10	HSSEX(1), HSAGEIR(1), DMARETHN(3), FEP(1), et BMPBMI(1)
3. Logistique	HAF10	HSSEX(1), HSAGEIR(1), DMARETHN(3), HAD1(1), HAE3(1), PBP(1), FEP(1), CHP(1), et G1P(1)
4 et 5. Linéaire	SPFVC	HSAGEIR(1), DMARETHN(3), HFA8R(2), et BMPBMI(1) [Selon le sexe (HSSEX), et limité aux personnes n'ayant jamais fumé (HAR1, HAR3)]
6 et 7. Logistique	HCHP (1 SI CHP>=240 ET 0 SINON)	HSAGEIR(2), DMARETHN(3), HFA8R(1), BMPBMI(3), (HAR3, HAR1)(2), BMPBMI*HSAGEIR(6), et DMARETHN*BMPBMI(9) [Selon le sexe (HSSEX)]
8. Logistique	HAC1E	HSAGEIR(5), HSSEX(1), (HAR3, HAR1)(2), SPPEAK(4), (HAR1, HAR3)(3), SPPEAK(3), et SPFVC(2) et SPFVC(1)

Biais standardisés simulés des estimateurs en présence d'imputation multiple pour les 115 coefficients de régression	
Fourchette des biais standardisés	Fréquence
-5,2	1
-1,5	1
-1,3	1
-1,1	1
(-1, -0,6]	2
(-0,6, -0,4]	2
(-0,4, -0,2]	3
(-0,2, 0)	52
(0, 0,2)	44
[0,2, 0,4)	6
[0,4, 0,6)	1
[0,6, 1)	1
3,7	115
Total	

d'imputation multiple ne sont que modérément plus efficaces que les estimateurs sans imputation, et que les estimateurs en présence d'imputation multiple sont affectés d'un certain biais, leurs erreurs quadratiques moyennes sont plus grandes que celles des estimateurs sans imputation dans 22 des 32 cas.

Les résultats des simulations en ce qui a trait à l'efficacité relative des estimateurs en présence d'imputation multiple concordent aussi avec la théorie. Puisque $V_{\text{comp}}^{VI}/V_{\text{SI}}^{VI}$ devrait être égal à environ 0,5, et que V_{IM}^{VI} devrait être un peu plus faible que V_{SI}^{VI} , il s'ensuit que $V_{\text{comp}}^{VI}/V_{\text{IM}}^{VI}$ devrait être légèrement supérieur à 0,5, ou de façon équivalente, que le ratio typique des écarts-types standardisés E_{-1}^{VI}/E_{-1}^{SI} devrait être un peu plus grand que $\sqrt{0,5} = 0,71$. En effet, la médiane des ratios résumés au tableau 4 est 0,75. Une alternative à l'estimation en présence d'imputation multiple est la pondération en deux phases basée sur les estimateurs des questions communes et leurs différences entre les blocs. Tout gain d'efficacité comparativement à l'estimation en présence d'imputation multiple serait dû à l'information supplémentaire provenant des questions échantillonnées.

3.3.2 Estimation des coefficients de régression

Nous avons également évalué les méthodes d'échantillonnage matriciel et d'imputation multiple dans le cas de l'estimation des coefficients de huit modèles de régression, que nous avons spécifiés de sorte qu'ils soient similaires aux modèles décrits dans la littérature. Les modèles de régression, qui sont énumérés au tableau 5, comportent, en tout, 115 coefficients. Les estimateurs sans imputation pour les coefficients de régression n'ont pas été inclus dans l'étude par simulation, mais nous discutons de certains résultats théoriques relatifs à leur efficacité à la présente section.

Pour chaque coefficient de régression, nous adoptons une définition du biais standardisé simulé analogue à celle utilisée pour chaque moyenne à la section 3.1. Le tableau 6 résume les biais standardisés pour les 115 coefficients de régression. La plupart de ces biais sont faibles, les valeurs absolues n'étant supérieures à 1 que pour cinq d'entre eux, et égales ou supérieures à 0,6 pour sept seulement. Le tableau 7 résume les ratios des écarts-types des estimations avec les données complètes sur les 100 ensembles de données simulés à ceux des estimations en présence d'imputation multiple, pour les 115 coefficients de régression. Des résumés distincts sont présentés selon que les modèles de régression contiennent des variables échantillonnées provenant d'un seul bloc (modèles 1, 2, 6 et 7) ou de deux blocs (modèles 3, 4, 5 et 8). Une plus forte

proportion de ratios s'approche de 1 que dans le cas de l'estimation des moyennes (tableau 4). En outre, pour plusieurs coefficients de régression (particulièrement provenant des modèles 3, 6, 7 et 8), les écarts-types simulés des estimateurs sur les données complètes sont modérément plus grands que ceux des estimateurs en présence d'imputation multiple, et pour un coefficient, le ratio est d'environ 2. Enfin, il existe quatre coefficients de régression pour lesquels il semble y avoir une perte importante d'efficacité due à l'échantillonnage matriciel, les ratios étant inférieurs à 0,3 (provenant, respectivement, des modèles 1, 2, 5 et 8). Les ratios proches de l'unité ou supérieurs à celle-ci pourraient être dus en partie à un mauvais ajustement de certains modèles de régression sur les données complètes et à un meilleur ajustement des modèles sur les données complètes par imputation, ce dernier résultant d'un processus d'imputation fondé sur des modèles de régression. En outre, les deux ratios les plus faibles sont observés pour des modèles de régression contenant des variables échantillonnées provenant de deux blocs, pour lesquelles la fraction de sujets dans l'échantillon matriciel sans données manquantes est seulement un sixième, comme nous en discutons plus loin.

Pour les modèles de régression comprenant des variables échantillonnées ne provenant que d'un seul bloc, l'efficacité théorique de l'estimateur sur données complètes relative-ment à l'estimateur sans imputation, c'est-à-dire le ratio de la variance du second à celle du premier, est approximativement égale à 2, parce qu'il n'existera des données complètes sur ces variables que pour approximativement la moitié des sujets compris dans l'échantillon matriciel; pour les modèles de régression contenant des variables échantillonnées provenant de deux blocs, l'efficacité théorique relative est d'environ 6. Par contre, les efficacités relatives simulées respectives de l'estimateur sur données complètes comparativement à l'estimateur en présence d'imputation multiple, c'est-à-dire les inverses des carrés des ratios résumés au tableau 7, sont inférieurs à 2 pour 64 des 69 coefficients lorsqu'un seul bloc de questions entre en jeu et ils sont inférieurs à 6 pour 44 des 46 coefficients lorsque deux blocs sont utilisés. Donc, les estimateurs en présence d'imputation multiple sont généralement plus efficaces que les estimateurs sans imputation pour les problèmes de régression. Néanmoins, les importantes pertes d'efficacité des estimateurs en présence d'imputation multiple relativement aux estimateurs sur données complètes observées pour certains coefficients, ainsi que les gains apparents d'efficacité pour d'autres coefficients justifieraient une étude plus poussée.

La perte d'efficacité due à l'échantillonnage matriciel plutôt qu'à l'utilisation d'un questionnaire complet peut être évaluée en comparant l'erreur d'échantillonnage des estimateurs en l'absence d'imputation, en présence d'imputation multiple et avec données complètes (calculée comme étant l'écart-type sur les 100 ensembles de données simulées). Le tableau 4 résume les ratios des écarts-types simulés des estimateurs en présence d'imputation multiple à ceux des estimateurs sans imputation, et les ratios des écarts-types simulés des estimateurs sur données complètes à ceux des estimateurs en présence d'imputation multiple (nous utilisons l'expression « écart-type simulé » d'un estimateur plutôt que « erreur-type simulée » pour éviter la confusion avec l'erreur-type estimée que l'on pourrait obtenir d'après l'analyse de chaque ensemble de données simulé).

Tableau 4		
Ratios des écarts-types simulés des estimateurs sans imputation (SI), en présence d'imputation multiple (IM) et sur les données complètes (comp) des moyennes de population pour les 32 questions échantillonnées		
Ratios	Fréquence	
$E-T_{IM}/E-T_{SI}$	$E-T_{comp}/E-T_{IM}$	
(0,5, 0,6]	2	2
(0,6, 0,7]	9	9
(0,7, 0,8]	14	14
(0,8, 0,9]	6	6
(0,9, 1]	1	1
(1, 1,03]	7	32
Total	32	32

Habituellement, les estimateurs en présence d'imputation multiple sont plus efficaces que les estimateurs sans imputation, mais le gain d'efficacité n'est que moyen, comme l'indique le fait que la plupart des ratios $E-T_{IM}/E-T_{SI}$ du tableau 4 sont compris entre 0,9 et 1. Des gains aussi modestes d'efficacité peuvent être prédits grossièrement d'après les indices de valeur prédictive basés sur les données provenant de la NHANES II (présentées au tableau 2), comme il suit. Puisque chaque question échantillonnée n'est incluse que dans la moitié des questionnaires d'échantillonnage matriciel, il s'ensuit que la variance d'un estimateur basé sur les données complètes de la moyenne d'une question échantillonnée devrait être égale à environ la moitié de la variance de l'estimateur sans imputation correspondant. Diviser le numérateur et le dénominateur de l'expression (3) par V_{SI} et fixer $V_{comp}/V_{SI} = 0,5$, donne $2(1 - V_{IM}/V_{SI})$ comme expression approximative de l'indice de valeur prédictive dans cette étude en simulation. Pour un indice de 0,12, qui est la médiane des indices « réajustés moyens » du tableau 2, il s'ensuit que V_{IM}/V_{SI} devrait être d'environ 0,94. Ce ratio des variances est équivalent à un ratio des écarts-types d'environ $\sqrt{0,94} = 0,97$, qui est proche du milieu de la fourchette des ratios résumés au tableau 4. Dans la présente étude, parce que les estimateurs en présence

l'estimateur sans imputation comme étant $(Moy_{SI} - Moy_{comp}) / E-T_{SI}$ ou $Moy_{SI} - Moy_{comp}$ et $E-T_{SI}$ de notent, respectivement, les moyennes des estimations sans imputation et sur les données complètes, ainsi que l'écart-type des estimations sans imputation sur l'ensemble des 100 ensembles de données simulées. Nous avons défini un biais standardisé simulé analogue pour l'estimateur en présence d'imputation multiple (IM). Le tableau 3 résume les biais standardisés simulés pour les 32 questions échantillonnées.

Tableau 3		
Biais standardisés simulés des estimateurs sans imputation et en présence d'imputation multiple des moyennes de population pour les 32 questions échantillonnées		
Fréquence	Sans imputation	Biais standardisé
1	1	-1,4
4	4	(-1, -0,6]
5	5	(-0,6, -0,4]
4	4	(-0,4, -0,2]
4	4	(-0,2, 0)
10	15	(0, 0,2)
4	17	[0,2, 0,4)
2	2	[0,4, 0,6)
1	1	[0,6, 1)
1	1	1,4
1	1	4,6
32	32	Total

Puisque notre mécanisme d'échantillonnage matriciel produit des données manquantes qui manquent entièrement au hasard, les estimateurs sans imputation sont presque sans biais, ce qui est reflété dans les résultats des simulations par le fait qu'aucun biais standardisé absolu n'est supérieur à 0,2. Les estimateurs en présence d'imputation multiple sont généralement affectés d'un biais standardisé simulé un peu plus élevé que les estimateurs sans imputation, quoique le biais standardisé absolu soit inférieur à 1 pour toutes les questions échantillonnées sauf trois et inférieur à 0,6 pour toutes, sauf sept. En guise de ligne directrice pour évaluer les biais standardisés, Cochran (1977, page 14) montre qu'un biais standardisé de 0,6 produit des intervalles de confiance à 95 % nominaux dont la couverture réelle est d'environ 91 %. Tout biais important observé dans le cadre de la présente étude lorsqu'on utilise l'échantillonnage matriciel conjugué à l'imputation multiple est vraisemblablement dû à des défauts dans les modèles d'imputation et non à l'échantillonnage matriciel proprement dit, puisque nous avons constaté que les analyses sans imputation étaient approximativement sans biais. En utilisant de plus grandes tailles d'échantillon dans une enquête à une enquête réelle, les biais standardisés correspondants auraient tendance à subir un mouvement à la hausse, à cause des dénominateurs plus faibles, mais ils pourraient aussi subir un mouvement à la baisse, à cause de l'amélioration des approximations sur grand échantillon.

était $w_{2ch} = 1$. Si la taille c dans la strate h était supérieure à 30, alors le poids d'échantillonnage de simulation de deuxième degré pour le sujet i dans la grappe c était $w_{2ch} \propto \pi_{1ch}^{-1}$, où π_{1ch} représentait la probabilité de sélection de premier tirage du sujet i . Pour chaque sujet échantillonné, le poids d'échantillonnage de simulation final était $w_{ich} = w_{1h} \times w_{2ch}$, $i = 1, 2, \dots, 30$, $c = 1, 2, h = 1, 2, \dots, 20$.

Les effets de plan pour l'estimation des moyennes de population étaient de l'ordre de 2,1, en moyenne, dans cette étude par simulation. Les caractéristiques du plan d'échantillonnage complexe utilisées dans l'étude sont informatives en ce sens qu'ignorer les caractéristiques du plan dans les analyses des données pourrait produire des estimations biaisées et une sous-estimation des variances d'échantillonnage. Cela est dû en particulier à l'utilisation de données sur la race, la pression artérielle et l'IMC dans le plan d'échantillonnage de simulation et au lien bien établi entre la race ou l'ethnicité et la pression artérielle ou l'IMC.

3.2.2 Simulation d'échantillons matriciels

Nous avons tiré 100 échantillons probabilistes indépendants à partir de la population finie. Chaque échantillon simulé comportait 1 200 sujets (20 strates de simulation, 2 grappes de simulation par strate de simulation, 30 sujets par grappe de simulation).

L'échantillonnage matriciel a été superposé à chaque échantillon simulé en affectant chacun des 1 200 sujets aléatoirement à l'un des six questionnaires contenant les questions communes et l'un des paires de blocs (1, 2), (1, 3), (1, 4), (2, 3), (2, 4) ou (3, 4). L'affectation aléatoire a été exécutée de façon que 200 sujets soient affectés à chaque questionnaire. Donc, pour chaque échantillon matriciel, les questions communes étaient disponibles pour l'ensemble des 1 200 sujets échantillonnés, tandis que chaque question échantillonnée était disponible pour 600 sujets échantillonnés.

3.2.3 Comparaison des méthodes d'estimation

Nous avons obtenu des estimations ponctuelles d'après chaque échantillon de l'étude en simulation selon trois méthodes, à savoir l'analyse des données complètes à titre d'étalon, l'analyse des données échantillonnées matriciellement sans imputation et l'application de l'imputation multiple pour remplacer les valeurs manquantes causées par l'échantillonnage matriciel, suivies de l'analyse des données multi-imputées. Pour l'analyse des données complètes et des données échantillonnées sans imputation, les estimations ponctuelles ont été pondérées. Pour les analyses en présence d'imputation multiple, nous avons utilisé les mêmes poids pour calculer l'estimation ponctuelle à partir de chacun des ensembles de données complètes par imputation multiple.

puis nous avons calculé la moyenne habituelle des estimations ponctuelles en présence d'imputation multiple (Rubin et Schenker 1986; Rubin 1987, section 3.1).

L'imputation multiple des valeurs des questions échantillonnées manquantes a été exécutée par la méthode de régression séquentielle (Kennickell 1991; Oudshoorn, Van Buuren et Van Rijckekevorssel 1999; Raghunathan, Lepkowski, Van Hoewyk et Solenberger 2001), telle qu'elle est implémentée dans le logiciel IVEware (<http://www.isr.umich.edu/src/smp/ive>). Nous avons créé cinq ensembles d'imputation indépendante les uns des autres en appliquant indépendamment la méthode de régression séquentielle cinq fois, avec dix itérations de l'algorithme de régression séquentielle pour chaque ensemble d'imputation. Le nombre d'imputations est basé sur la théorie et l'expérience indiquant que cinq imputations sont habituellement suffisantes, surtout si la fraction d'information manquante n'est pas importante (Rubin 1996). Pour des taux de données manquantes de 50 % pour les questions échantillonnées, la fraction d'information manquante, qui est approximativement $1 - V_{comp}^{imp} / V_{imp}$, devrait, en principe, être au plus de 50 %, ce que confirment les résultats des simulations. Selon Rubin (1987, tableau 4.1), l'efficacité relative de cinq imputations sur grand échantillon comparativement à un nombre infini d'imputations est de 90 % lorsque 50 % d'information manque. Un grand nombre d'imputations augmenterait la précision de l'estimation de la variance entre imputations (V_{imp}^{imp}) et la fraction d'information manquante.

Pour tenir compte du plan d'échantillonnage simulé complexe, les effets principaux ont été inclus dans le modèle d'imputation pour la strate de simulation et la grappe de simulation emboîtée dans la strate de simulation. Le logarithme du poids d'échantillonnage de simulation a également été inclus comme prédicteur dans le modèle d'imputation, ainsi que les questions communes et les questions échantillonnées.

3.3 Résultats de l'étude par simulation

Pour évaluer les estimations basées sur le plan d'échantillonnage matriciel, nous avons considéré deux types de problèmes analytiques, à savoir l'estimation des moyennes de population des questions échantillonnées et les analyses par la régression faisant intervenir les questions échantillonnées et les questions communes. Nous avons comparé les propriétés des estimateurs sans imputation, en présence d'imputation multiple et pour les données complètes sur l'ensemble des 100 ensembles de données simulées, afin d'évaluer le biais et la perte d'efficacité dus à l'échantillonnage matriciel conjugué à l'imputation multiple.

3.3.1 Estimation des moyennes de population pour les questions échantillonnées

Pour la moyenne de population d'une question échantillonnée, nous avons défini le biais standardisé simulé de

3.2 Conception de l'étude par simulation fondée sur

les données de la NHANES III

3.2.1 Population et plan d'échantillonnage

Le plan d'échantillonnage matriciel et l'analyse en présence d'imputation multiple pourraient être appliqués à l'échantillon complet de la NHANES III. Cela serait certes informatif, mais une étude basée sur un seul ensemble de données ne permettrait pas d'évaluer les propriétés statistiques des méthodes étudiées sous échantillonnage répété. Par conséquent, les 11 759 sujets de la NHANES III qui avaient fourni des données complètes sur les variables énumérées au tableau 1 ont été traités comme une population finie et des échantillons répétés ont été tirés à partir de cette population. Nous avons utilisé pour sélectionner les échantillons un plan d'échantillonnage complexe plutôt qu'un échantillonnage aléatoire simple afin de créer une étude par simulation plus réaliste. Pour réaliser cet objectif, nous avons ajouté trois variables de plan d'échantillonnage à la population finie, à savoir 1) strate de simulation, 2) grappe de simulation et 3) poids d'échantillonnage de simulation (ici, le qualificatif « de simulation » est utilisé pour faire la distinction entre ces quantités et les variables du plan d'échantillonnage original de la NHANES III).

1. **Strates de simulation :** L'échantillon du fichier de données à grande diffusion de la NHANES III comporte 49 strates contenant chacune 2 grappes. La stratégie, pour l'étude en simulation, consistait à créer un plus petit nombre de strates contenant chacune un plus grand nombre de grappes, pour s'assurer que la variation d'un échantillon à l'autre entre les échantillons simulés soit suffisante. Nous avons regroupé les 49 strates originales en 20 strates de simulation de la façon suivante. Chacune des 49 strates originales a été classée dans l'une de huit catégories formées par recoupement de la région de recensement (quatre niveaux) et de la situation de région rurale/urbaine basée sur le code du United States Department of Agriculture (deux niveaux). Dans chacune de ces huit catégories, nous avons procédé à une analyse typologique en utilisant les proportions de non-Blancs au niveau de la strate pour sélectionner les strates originales qu'il convenait de combiner. La combinaison des strates originales a créé de deux ou trois strates de simulation dans chacune des huit catégories, ce qui a donné en tout 20 strates de simulation. Cette méthode de création de strates plus grandes a également augmenté l'hétérogénéité raciale entre les strates de simulation résultantes, ce qui a accru l'importance de la pondération dans les analyses.

2.

Grappes de simulation : L'échantillon du fichier de données à grande diffusion de la NHANES III comporte 98 grappes, c'est-à-dire 2 grappes dans chacune des 49 strates originales. Après avoir regroupé les 49 strates originales en 20 strates de simulation, nous avons réparti les grappes originales en 30 grappes d'échantillonnage de simulation. Le nombre de grappes de simulation par strate de simulation variait de 3 à 25, et le nombre de sujets par grappe de simulation variait de 30 à 98.

3.

Poids d'échantillonnage de simulation : Nous avons déterminé les poids d'échantillonnage de simulation au moyen du plan d'échantillonnage à deux degrés suivant. Premièrement, pour chaque strate de simulation, nous avons tiré deux grappes de simulation par échantillonnage aléatoire simple sans remise. Comme les nombres de grappes de simulation dans les 20 strates de simulation étaient inégaux, le poids d'échantillonnage de simulation correspondant à cette étape était $w_h = A_h / 2$, $h = 1, 2, \dots, 20$, où A_h est le nombre de grappes de simulation dans la strate de simulation h . Deuxièmement, à partir de chaque grappe de simulation sélectionnée, nous avons tiré 30 sujets au hasard sans remise avec des probabilités de sélection variables. Si la taille de la grappe était égale à 30, alors tous les sujets ont été inclus dans l'échantillon. Pour les grappes contenant plus de 30 sujets, nous avons calculé les probabilités de sélection de premier tirage en normalisant les inverses des poids originaux provenant de l'échantillon du fichier de données à grande diffusion de la NHANES III de sorte que leur somme soit égale à 1 dans chaque grappe de simulation, l'inverse normalisé étant utilisé pour chaque sujet comme probabilité de sélection de ce sujet. Les probabilités de sélection de premier tirage dans les grappes de simulation variaient de 0,0003 à 0,2756. Soit i l'indice représentant les sujets échantillonnés dans une grappe de simulation, c les grappes échantillonnées dans une strate de simulation et h les strates de simulation telles qu'elles sont décrites plus haut, $i = 1, 2, \dots, 30$, $c = 1, 2, \dots, 20$. Si la taille de la grappe c dans la strate h était égale à 30, alors le poids d'échantillonnage de simulation de deuxième degré pour le sujet i dans la grappe c

Tableau 1 (suite)

Variables de la NHANES III qui ont été incluses dans l'évaluation.
Les questions marquées d'un astérisque ne figuraient pas dans la NHANES II

Nom de la variable	Description de la variable	Type
BMPWHR*	Ratio tour de taille-tour de hanche	Échantillonnée – 2
HACIE	Un docteur vous a-t-il déjà dit que vous aviez : asthme	Échantillonnée – 2
HACIK	Un docteur vous a-t-il déjà dit que vous aviez : maladie thyroïdienne	Échantillonnée – 2
HAF24	Engourdissement, etc., d'un côté du visage/corps pendant plus de 5 minutes	Échantillonnée – 2
HALI1B	Yeux larmoyants, qui chatouillaient au cours des 12 derniers mois	Échantillonnée – 2
HALI9A*	Au cours des 12 derniers mois, a eu : rhume ou grippe	Échantillonnée – 2
HALI9C*	Au cours des 12 derniers mois, a eu : pneumonie	Échantillonnée – 2
HAT28	Actif(ve) comparativement aux hommes/femmes de votre âge	Échantillonnée – 2
PBP	Plomb (ug/dL)	Échantillonnée – 2
SPFVC*	CVF, valeur la plus grande (mL)	Échantillonnée – 2
FEP	Fer sérique (ug/dL)	Échantillonnée – 3
HAF1	A déjà eu une douleur ou une gêne dans la poitrine	Échantillonnée – 3
HAF23	Faiblesse/paralysie dans le visage, le bras, la jambe pendant plus de 5 minutes	Échantillonnée – 3
HAF19B	Au cours des 12 derniers mois, a eu : sinusite/problème de sinus	Échantillonnée – 3
HAR1	Avez-vous fumé 100 cigarettes ou plus au cours de votre vie	Échantillonnée – 3
HAR3	Fumez-vous des cigarettes à l'heure actuelle	Échantillonnée – 3
SPPEAK*	Débit maximal expiratoire	Échantillonnée – 3
BDPBOMB*	Densité minérale osseuse, région totale (g/cm ³)	Échantillonnée – 4
HAB4	Au cours des 12 derniers mois, nombre d'hospitalisations	Échantillonnée – 4
HAC1D	Un docteur vous a-t-il déjà dit que vous aviez : accident vasculaire cérébral	Échantillonnée – 4
HAC1F	Un docteur vous a-t-il déjà dit que vous aviez : bronchite chronique	Échantillonnée – 4
HAC1H	Un docteur vous a-t-il déjà dit que vous aviez : rhume des foies	Échantillonnée – 4
HAC1I	Un docteur vous a-t-il déjà dit que vous aviez : cataracte	Échantillonnée – 4
HAF6*	A-t-on déjà vérifié votre cholestérol sanguin	Échantillonnée – 4
HAMI1*	Considère que son poids est excessif/insuffisant/correct	Échantillonnée – 4
HAF7*	Un docteur a dit que le taux de cholestérol sanguin était élevé	Échantillonnée – 4

Tableau 2
Indices de valeur prédictive basés sur les données de la NHANES II pour les questions échantillonnées dans le plan d'échantillonnage matriciel

Question	Bloc	Optimaux	Réalisés
HAC1J(GOITRE)	1	0,04	0,15
HAC1O(AUTRE CANCER)	1	0,05	0,06
HALI1A(SYMPÔMES NASAUX)	1	0,17	0,27
G1P(GLUCOSE SÉRQUE)	1	0,26	0,30
HAC1E(ASTHME)	2	0,09	0,09
HAC1K(MALADIE THYROÏDIENNE)	2	0,07	0,07
HAF24(ENGOURDISSEMENT)	2	0,12	0,11
HALI1B(YEUX LARMOYANTS)	2	0,14	0,15
HAT28(ACTIF(VE) POUR SON ÂGE)	2	0,12	0,13
PBP(PLOMB (ug/dL))	2	0,19	0,20
HAF1(DOULEUR DANS LA POITRINE)	3	0,25	0,25
HAF23(FAIBLESSE/PARALYSIE)	3	0,08	0,12
HAL19B(SINUSITE/SINUS)	3	0,07	0,12
HAR1(100 CIGARETTES OU PLUS)	3	0,13	0,14
HAR3(FUME À L'HEURE ACTUELLE)	3	0,10	0,11
FEP(FER SÉRQUE)	3	0,05	0,05
HAB4(NOMBRE D'HOSPITALISATIONS)	4	0,07	0,11
HAC1D(ACCIDENT VASCULAIRE CÉRÉBRAL)	4	0,19	0,20
HAC1F(BRONCHITE)	4	0,10	0,12
HAC1H(RHUME DES FOIES)	4	0,07	0,09
HAC1J(CATARRACTE)	4	0,08	0,12

Nota : Les indices prédictifs optimaux sont déterminés pour chaque question séparément et ne sont pas nécessairement réalisables pour toutes les questions simultanément.

élevées de l'indice effectivement réalisées au moyen du plan sélectionné dans les trois blocs ne contenant pas la question en question (« Réalisés »). Les valeurs de l'indice sont très élevées de la plus faible à la plus élevée pour chaque question considérée, de sorte que les colonnes du tableau contenant les valeurs d'indice ne correspondent pas à des questions ni à des blocs particuliers. Le tableau 2 montre que le plan choisi est presque optimal pour le critère de la section 2.2.1. Par exemple, l'écart moyen entre les indices prédicatifs optimaux et les indices correspondants effectivement réalisés n'est que de 0,002.

La colonne du tableau 2 étiquetée « Inférieur » sous « Réalisés » donne les bornes inférieures sur l'amélioration attendue des estimateurs des moyennes univariées pour les questions échantillonnées. Dix-neuf des 21 indices prédicatifs figurant dans cette colonne sont inférieurs à 0,20, ce qui donne à penser que l'efficacité des estimateurs multi-imputés est relativement faible dans ce plan d'échantillonnage matriciel. Pour une discussion plus approfondie de cette question, voir les sections 3.3 et 4.

Tableau 1

Variables de la NHANES III qui ont été incluses dans l'évaluation.
Les questions marquées d'un astérisque ne figuraient pas dans la NHANES II

Nom de la variable	Description de la variable	Type
BMPBMI	Indice de masse corporelle	Commune
CHP*	Cholestérol sérique (mg/dL)	Commune
DMARETHN	Race-ethnité	Commune
DMPFCRGN	Région de recensement, pondération (Texas dans le Sud)	Commune
DMPMETRO	Code de région rurale/urbaine basé sur le code Usda	Commune
GHP*	Hémoglobine glycosylée (%)	Commune
HAB1	Votre santé est-elle, en général, excellente, ..., mauvaise	Commune
HAB2*	Se rend dans un endroit particulier pour obtenir des soins de santé	Commune
HAB5*	Au cours des 12 derniers mois, nombre de visites chez un docteur	Commune
HAC1C	Un docteur a déclaré : insuffisance cardiaque congestive	Commune
HAC1L*	Un docteur vous a-t-il déjà dit que vous aviez : l'insus	Commune
HAC1M	Un docteur vous a-t-il déjà dit que vous aviez : goutte	Commune
HAD1	Vous a-t-on déjà dit que vous aviez du sucre/diabète	Commune
HAD10	Pour le moment, prenez-vous des pilules contre le diabète	Commune
HAB3	On vous a dit au moins deux fois que vous faisiez de l'hypertension	Commune
HAF10	Un docteur vous a-t-il déjà dit que vous aviez fait une crise cardiaque	Commune
HAF26	Étourdissement grave pendant plus de 5 minutes	Commune
HAT1	Toux la plupart des jours, 3 mois consécutifs ou plus dans l'année	Commune
HAL6	A eu des sifflements dans la poitrine au cours des 12 derniers mois	Commune
HAL14E	Symptômes déclenchés par le pollen	Commune
HAZMNK1R	Pa K1 moyenne d'après questionnaire-ménage et CEM	Commune
HAZMNK5R	Pa K5 moyenne d'après questionnaire-ménage et CEM	Commune
HFA12	État matrimonial	Commune
HFA8R	Grade le plus élevé ou nombre d'années d'école achevées	Commune
HSAG6IR	Âge au moment de l'entrevue (questionnaire de présélection)	Commune
HSEX	Sexe	Commune
IIP	Insuline sérique (uU/mL)	Commune
GIP	Glucose sérique (mg/dL)	Echantillonnée – 1
HAC1J	Un docteur vous a-t-il déjà dit que vous aviez : goitre	Echantillonnée – 1
HAC1N*	Un docteur vous a-t-il déjà dit que vous aviez : cancer de la peau	Echantillonnée – 1
HAC1O	Un docteur vous a-t-il déjà dit que vous aviez : autre cancer	Echantillonnée – 1
HAF14*	Ressent de la douleur dans les deux jambes en marchant	Echantillonnée – 1
HAL11A	Lourd, chatouillement ou écoulement nasal au cours des 12 derniers mois	Echantillonnée – 1

2.2.2 Un algorithme d'affectation

Les critères de la section 2.2.1 doivent être maximisés sur un ensemble d'entrées entières (affectations aux blocs) d'une fonction, sous un ensemble de contraintes linéaires imposées par la nécessité de créer des blocs de longueur approximativement égale avec des questions éventuellement reliées. Bien que les méthodes de programmation entières puissent être appliquées à cette maximisation, l'algorithme qui suit est nettement plus simple et donne des résultats presque optimaux pour l'application de la NHANES, comme nous le démontrons à la section 3.1.

Étape 1. Ordonner aléatoirement les questions échantillonées. L'affectation des questions aux blocs se fait séquentiellement, par répétition des étapes 2 et 3 qui suivent, jusqu'à ce que toutes les questions aient été affectées.

Étape 2. Affecter la prochaine (ou la première) question non affectée, disons $Y^{(0)}$, au bloc contenant le plus petit nombre de questions. Si plusieurs blocs sont à égalité, affecter $Y^{(0)}$ à celui pour lequel l'indice prédictif maximum $I(Y^{(0)} | X, Z)$ est le plus faible pour $Y^{(0)}$. S'il persiste un ex aequo, affecter $Y^{(0)}$ à n'importe lequel des blocs admissibles. S'il existe des questions liées à $Y^{(0)}$, les affecter également au bloc sélectionné.

Étape 3. Pour chaque question affectée à l'étape 2 ($Y^{(0)}$ ou les questions qui y sont liées), trouver les questions non affectées restantes, disons $Y^{(1)}$, qui en sont les plus prédictives. Affecter $Y^{(1)}$ (et toute question liée à $Y^{(1)}$) à un autre bloc que celui sélectionné à l'étape 2, en suivant la même procédure que celle utilisée pour $Y^{(0)}$ à l'étape 2. L'expérience avec les données de la NHANES donne à penser que l'algorithme est moyennement sensible au classement initial des questions (à l'étape 1). Pour réduire la dépendance, nous avons généré 1 000 plans d'expérience avec classements sélectionnés aléatoirement, puis nous avons choisi celui donnant la meilleure mesure globale de la valeur prédictive (telle qu'elle est définie à la section 2.2.1).

3. Étude au moyen de données provenant de la NHANES

Pour évaluer la faisabilité d'un plan d'échantillonnage matriciel pour une enquête telle que la NHANES, nous avons réalisé une étude d'évaluation. Pour commencer, nous avons utilisé la NHANES II (c'est-à-dire, la deuxième NHANES) pour créer un plan d'échantillonnage matriciel par la méthode décrite à la section 2. Ce plan simule la situation réelle dans laquelle les données provenant d'une enquête antérieure sont utilisées pour concevoir le question-naire d'une nouvelle enquête. Puis, nous avons appliqué le plan ainsi établi à plusieurs échantillons simulés créés d'après les données de la NHANES III. Les participants à la

NHANES III pour lesquels existaient des données complètes pour un ensemble sélectionné de variables ont été traités comme une grande population finie. Nous avons tiré 100 échantillons à partir de la population finie de la NHANES III selon un plan d'échantillonnage stratifié à deux degrés avec probabilités de sélection inégales. Les données complètes étant disponibles pour chacun des échantillons simulés, ceux-ci constituent notre « étalon ». Nous avons ensuite imposé le plan d'échantillonnage matriciel à chaque échantillon et procédé à l'imputation multiple des valeurs manquantes dues à l'échantillonnage matriciel. Enfin, nous avons réalisé plusieurs analyses en utilisant les échantillons matriciels sans imputation, les échantillons et les échantillons avec données complètes (c'est-à-dire l'étalon). Les résultats résumés sur l'ensemble des échantillons simulés produisent des estimations des propriétés des diverses méthodes sous échantillonnage répété. Le plan d'échantillonnage matriciel créé en utilisant les données de la NHANES II est résumé à la section 3.1. Le plan de l'étude par simulation au moyen des données de la NHANES III est décrit à la section 3.2. Les résultats de l'étude sont présentés à la section 3.3. Certains limites de l'étude qui ne sont pas abordées aux sections 3.1 à 3.3 sont décrites à la section 3.4.

3.1 Plan d'échantillonnage matriciel fondé sur les données d'apprentissage provenant de la NHANES II

Étant donné le temps qu'aurait pris l'extraction et l'analyse de toutes les variables de la NHANES III, nous n'avons inclus qu'un sous-ensemble dans l'étude afin que celle-ci demeure gérable, quoique le logiciel utilisé permettrait de traiter un beaucoup plus grand nombre de variables. Les variables retenues pour l'étude comprennent les questions représentant nombre des sujets couverts dans l'enquête et ont été sélectionnées en consultation avec les spécialistes du domaine. Les types de données comprennent des variables binaires et continues représentant des questions d'enquête et des mesures de laboratoire. Une paire de questions formant un enchaînement a été incluse : « Avez-vous fumé 100 cigarettes ou plus au cours de votre vie? » suivi de « Fumez-vous à l'heure actuelle? » L'algorithme utilisé pour affecter les questions échantillonées aux blocs, décrit à la section 2.2, a forcé ces deux questions à être dans le même bloc.

Le tableau 1 donne une brève description des variables incluses. Les variables qui figuraient dans la NHANES III mais non dans la NHANES II (de nouveau, une situation réaliste) sont indiquées par un astérisque à côté de leur nom. Comme nous l'avons mentionné plus haut, le plan d'échantillonnage matriciel a été construit avec quatre blocs.

ce qui implique que les dérivées $\hat{\tau}_{m_{ms},j}^{ms}$ (β) sont égales aux moyennes des questions communes X et de la question échantillonnée Z .
Maintenant, représentons par $\hat{\tau}_{m_{ms}}^{ms}$ (β) le vecteur des dérivées et par $\tau_{m_{ms}}^{ms}$ la proportion de sujets pour lesquels des valeurs de X manquent. En appliquant l'équation (10) de Schaffer et Schenker (2000), avec leur fonction g égale à l'identité, et leur paramètre général θ égal à β , nous obtenons

$$V_{imp} \approx P^2 \left\{ m_{m_{ms}}^{ms} \sum_{i=1}^{m_{m_{ms}}^{ms}} \pi((X_T^i, Z_i) | \beta) [1 - \pi((X_T^i, Z_i) | \beta)] + (\hat{\tau}_{m_{ms}}^{ms}(\beta))^T V_{obs}^{ms}(\beta) \hat{\tau}_{m_{ms}}^{ms}(\beta) \right\} \quad (4)$$

si X est binaire, et

$$V_{imp} \approx P^2 [\sigma^2 / n_{ms} + (\hat{\tau}_{m_{ms}}^{ms}(\beta))^T V_{obs}^{ms}(\beta) \hat{\tau}_{m_{ms}}^{ms}(\beta)] \quad (5)$$

2.1.6 Estimation de l'indice de valeur prédictive d'après un échantillon d'apprentissage

Puisque nous supposons que les données manquantes prévues dans nos plans d'échantillonnage matriciel

et les estimations d'autres paramètres provenant d'un échantillon d'apprentissage peuvent être utilisées pour estimer les moments et paramètres correspondants dans les sous-échantillons contenant des valeurs observées de X , ainsi que dans ceux contenant des valeurs manquantes de X , sous l'hypothèse que l'échantillon d'apprentissage est tiré à partir de la même population cible. Les moments et paramètres incluent : $V(X)$, la variance résiduelle σ^2 , qui peut être estimée par $\hat{\sigma}^2$, la variance résiduelle estimée d'après la régression ajustée à l'échantillon d'apprentissage; les coefficients de régression, avec les estimations $\hat{\beta}^{ms}$ provenant de l'échantillon d'apprentissage; et la matrice de covariance des coefficients de régression, qui peuvent être approximés par rééchantillonnement de l'estimation $V^{tt}(\hat{\beta}^{tt})$ provenant de l'échantillon d'apprentissage pour obtenir $V_{obs}^{ms}(\hat{\beta}) \approx (n^{tt}/n^{obs}) V^{tt}(\hat{\beta}^{tt})$, où n^{tt} est la taille de l'échantillon d'apprentissage. Les dérivées $\hat{\tau}_{m_{ms}}^{ms}(\beta)$ et la fonction faisant intervenir μ dans (4) sont également sous la forme de moyennes de sous-échantillon et peuvent donc être estimées par les moyennes correspondantes dans l'échantillon d'apprentissage. En dénotant les dérivées dans l'échantillon d'apprentissage par $\hat{\tau}_{m_{ms}}^{tt}(\hat{\beta}^{tt})$, et en introduisant les estimateurs provenant de l'échantillon d'apprentissage par substitution dans (4) et (5), nous obtenons

$$V_{imp}^{mp} = P^2 (\hat{\sigma}_z^2 / n_{m_{ms}}^{ms} + \hat{\sigma}_z^2 / n_{obs}^{ms}), \quad (7)$$

pour les variables binaires, et

$$V_{imp}^{mp} \approx P^2 \left\{ \sum_{i=1}^{m_{m_{ms}}^{ms}} \left(n_{m_{ms}}^{tt} \pi((X_T^i, Z_i) | \hat{\beta}^{tt}) [1 - \pi((X_T^i, Z_i) | \hat{\beta}^{tt})] \right) + \frac{m_{m_{ms}}^{obs}}{n^{ms}} (\hat{\tau}_{m_{ms}}^{tt}(\hat{\beta}^{tt}))^T V^{tt}(\hat{\beta}^{tt}) \hat{\tau}_{m_{ms}}^{tt}(\hat{\beta}^{tt}) \right\} \quad (6)$$

2.2 Affectation des questions échantillonnées aux blocs

2.2.1 Critères de conception des questionnaires

Les questionnaires d'échantillonnage matriciel sont créés en affectant les questions échantillonnées à divers blocs, comme il est décrit au début de la section 2. Quatre objectifs de conception orientent l'affectation des questions : 1) affecter chaque question échantillonnée à un seul bloc; 2) affecter un nombre approximativement égal de questions à chaque bloc; 3) affecter à un même bloc les questions reliées logiquement et 4) affecter à chaque bloc une ou plusieurs questions qui prédisent les questions omises dans le bloc. Dénotons le nombre de blocs par n^{block} ($n^{block} = 4$ dans l'étude par simulation de la NHANES).

Un critère quantitatif pour le quatrième objectif est spécifié séparément pour chaque question échantillonnée X en trouvant les ($n^{block} - 1$) autres questions échantillonnées Z possédant les valeurs d'indice prédictif $I(X|X, Z)$ les plus élevées, pour l'affectation éventuelle de l'une d'entre elles aux ($n^{block} - 1$) blocs ne contenant pas X . Les questions Z ne comprennent pas celles liées à la question X , qui doivent figurer avec cette dernière dans un bloc. Les ($n^{block} - 1$) valeurs de $I(X|X, Z)$ pour les questions Z fournissent une borne supérieure sur les indices prédictifs qui pourraient être réalisés pour X . Comme ces valeurs d'indice optimales sont déterminées séparément pour chaque question échantillonnée X , elles pourraient ne pas être réalisables simultanément pour toutes les questions X . Pour évaluer un plan d'échantillonnage matriciel partiel, nous déterminons la valeur de l'indice $I(X|X, Z)$ la plus élevée effectivement obtenue pour chacun des ($n^{block} - 1$) blocs ne contenant pas de question échantillonnée X . Puis, nous calculons la moyenne des ($n^{block} - 1$) écarts entre ces indices et les indices prédictifs optimaux correspondants pour X . Enfin, nous calculons la moyenne de ces écarts moyens sur l'ensemble des questions échantillonnées X pour obtenir une mesure globale pour le plan.

$$V_{\text{imp}} = \lim_{M \rightarrow \infty} M^{-1} \sum_{j=1}^f (Y_j - \bar{Y})^2. \quad (2)$$

Nous supposons tout au long de l'exposé que le modèle d'imputation est compatible avec le modèle à données complètes, de sorte que l'estimateur de la variance (2) est convergent (Rubin 1987, section 3.6; Meng 1994).

2.1.4 Définition de l'indice

Lorsque des données sont recueillies auprès d'échantillons matriciels, il est possible d'obtenir des estimateurs simples, mais éventuellement inefficaces, de valeurs som-maires univariées d'une question échantillonnée, X , à partir des données observées sans aucune imputation (autrement dit, en utilisant uniquement les valeurs observées de X), parce que les valeurs manquantes de X manquent entièrement au hasard, la variance de l'estimateur sans imputation de $E(X)$ est dénotée $V_{\text{SI}} = V(X)/n_{\text{obs}}$.

L'indice proposé est la proportion de l'écart entre V_{SI} et V_{comp} qui est recouverte par l'estimateur sous imputation multiple, dans lequel est intégrée l'information contenue dans X et Z :

$$I(X|X, Z) = \frac{V_{\text{SI}} - V_{\text{IM}}}{V_{\text{SI}} - V_{\text{comp}}}. \quad (3)$$

(C'est-à-dire $V_{\text{IM}} = V_{\text{SI}}$.) L'indice peut être utilisé pour évaluer la contribution éventuelle de chaque question échantillonnée Z à l'estimation de la moyenne de chaque autre question échantillonnée X . Un plan d'échantillonnage matriciel désirable assure que pour chaque question échantillonnée X qui est exclue d'un bloc, il existe dans ce dernier d'autres questions échantillonnées Z dont l'indice de valeur prédictive de X est élevé, de sorte que l'information sur X peut être recouverte durant les analyses des données provenant de l'échantillon matriciel.

Nota :

1. Les variances V_{SI} , V_{comp} et V_{imp} sont proportionnelles à n_{tot}^{-1} , donc $I(X|X, Z)$ est indépendant de n_{tot} .

2. Si les questions communes X sont fortement prédictives de Z , l'indice ne fera pas de grande distinction entre les questions échantillonnées restantes Z ; mais, dans cette situation, la sélection de la question échantillonnée Z appropriée pour prédire X est

moins importante, puisque X est déjà bien prédite par X .

2.1.5 Approximation de V_{imp}

Pour faciliter le calcul de l'indice $I(X|X, Z)$, il est utile d'obtenir une approximation de la variance V_{imp} . Celle qui est développée ici a trait à un plan d'échantillonnage matriciel particulier, en supposant qu'un plan ait été choisi. Supposons que la loi de X sachant (X, Z) suit un modèle linéaire généralisé avec une fonction de lien μ qui dépend des paramètres inconnus β ,

$$E(X|X, Z) = \mu((X^T, Z)^T, \beta),$$

où la fonction de lien est égale à l'identité pour la variable continue X , $\mu(X) = X$, et à la fonction logarithmique pour la variable binaire X , $\mu(X) = \logit^{-1}(X)$. Si la variable X est continue, nous supposons aussi que la variance résiduelle, σ^2 , est constante. Bien qu'elles ne le soient pas ici, des extensions de ces modèles et méthodes peuvent être élaborées pour des variables catégoriques ou catégoriques ordonnées. Les catégories individuelles peuvent être représentées par des variables binaires, ou peuvent être regroupées en catégories sommatoires lorsqu'elles sont nominales. Schaffer et Schenker (2000) ont dérivé une approximation de la variance entre les ensembles de données imputés, c'est-à-dire V_{imp} , quand l'estimation calculée pour chaque ensemble de données complètes est une fonction lisse g des moyennes des variables concernées. (Dans le cas qui nous occupe, g est l'identité.) Cette approximation, qui est basée sur des développements en série de Taylor de premier ordre de g et μ , ainsi que sur des résultats sur grand échantillon tirés de la théorie des sondages (par exemple, Wolter 1985, chapitre 6), est celle que nous utiliserons ici.

L'estimation du maximum de vraisemblance (ou de quasi-vraisemblance) (McCullagh et Nelder 1989) de β fondée sur les n_{obs} sujets pour lesquels des valeurs de X sont observées donne un estimateur, $\hat{\beta}$, dont la matrice de variance-covariance est $V_{\text{obs}}(\hat{\beta})$ (rappelons l'hypothèse simplifiée 4 de la section 2.1.2). Soit $\hat{\mu}_{\text{SI}}(\hat{\beta}) \equiv \mu_{\text{SI}}(\hat{\beta})$, et dénotons sa dérivée par rapport à la j^{e} composante de $\hat{\beta}$ évaluée à $\hat{\beta}$ par $\hat{\mu}_{\text{SI},j}(\hat{\beta})$, $j = 1, \dots, (c+1)$. La dérivée est de la forme

$$\hat{\mu}_{\text{SI},j}^{\text{ms},f}(\hat{\beta}) = n^{-1} \sum_{i=1}^{n_{\text{ms}}} X_{ij}^f \hat{\mu}_{\text{SI},j}(\hat{\beta}, X_i^T, Z_i^T) \quad j = 1, \dots, c$$

et

$$\hat{\mu}_{\text{SI},j}^{\text{ms},c+1}(\hat{\beta}) = n^{-1} \sum_{i=1}^{n_{\text{ms}}} Z_i^j \hat{\mu}_{\text{SI},j}(\hat{\beta}, X_i^T, Z_i^T),$$

où $\hat{\mu}_{\text{SI}}(\hat{\beta}, X_i^T, Z_i^T) = \mu((X_i^T, Z_i^T)^T | 1 - \mu((X_i^T, Z_i^T)^T | \hat{\beta}))$ quand X est binaire. Si X est continue, $\hat{\mu}_{\text{SI}}(\hat{\beta}, X_i^T, Z_i^T) = 1$,

Comme nous l'avons mentionné plus haut, un plan d'échantillonnage matriciel crée ce que l'on peut considérer comme des données manquantes. Donc, dans un plan d'échantillonnage matriciel possible, les sujets doivent être classés de telle sorte que les n_{obs} sujets pour lesquels il existe des valeurs observées de X soient énumérés pour commencer, les valeurs de X étant dénotées par $X_1, \dots, X_{n_{\text{obs}}}$, et que les n_{mis} sujets pour lesquels les valeurs de X manquent suivent, les valeurs de X étant dénotées par $X_{n_{\text{obs}}+1}, \dots, X_{n_{\text{tot}}}$, où $n_{\text{tot}} = n_{\text{obs}} + n_{\text{mis}}$ est le nombre total d'observations.

L'espérance et la variance de X dans la population visée par l'échantillonnage matriciel sont dénotées par $E(X)$ et $V(X)$.

Plusieurs hypothèses sont faites pour simplifier le calcul des plans d'échantillonnage matriciel possibles.

1. Chaque prédicteur échantillonné Z est considéré individuellement lors de l'ajout aux questions communes X . S'il existe plusieurs questions présentant une forte corrélation mutuelle, l'algorithme d'affectation s'efforce d'attribuer ces questions à des blocs différents, comme le requiert un plan matriciel efficace. Une approche multivariée basée sur les corrélations partielles tenant compte d'autres questions échantillonnées produirait, en principe, des propriétés semblables, mais nécessiterait un beaucoup plus grand nombre de calculs.

2. Nous supposons que chaque prédicteur échantillonné Z est entièrement observé, alors qu'en pratique, il ne sera pas toujours possible de prédire X , parce que Z est aussi une question échantillonnée. En outre, l'occurrence de données manquantes non prévues (c'est-à-dire, des données manquantes qui ne sont pas créées par l'échantillonnage matriciel) n'est pas prise en compte. Bien que ces hypothèses puissent donner lieu à une surestimation de l'utilité de Z pour l'amélioration des estimations de $E(X)$, ce genre de surestimation peut être corrigé à l'aide de méthodes multivariées utilisant plusieurs variables Z . De surcroît, chaque question Z sera posée le même nombre de fois, de sorte que tout biais systématique dans la valeur prédictive devrait être approximativement le même pour chaque question échantillonnée; l'indice est utilisé principalement pour

établir le classement des questions, lequel ne sera pas modifié par un biais commun.

3. Pour le calcul des valeurs de l'indice, nous supposons que les répondants ont été sélectionnés par échantillonnage aléatoire simple dans l'échantillon matriciel ainsi que dans l'échantillon d'apprentissage. De nouveau, nous ne pensons pas qu'une surestimation cohérente de la précision réduira considérablement la performance de l'indice.

4. Nous supposons que l'échantillonnage matriciel produit des données manquantes selon le mécanisme de données manquant entièrement au hasard. Cette hypothèse est satisfaisante pour tous les plans d'échantillonnage matriciel pris en considération.

5. En vue de dériver les approximations qui suivent, nous supposons que n_{tot} est grand et que le ratio $n_{\text{obs}}/n_{\text{tot}}$ est constant quand n_{tot} augmente. Cette approximation devrait être adéquate pour la plupart des paramètres à estimer dans les enquêtes nationales.

2.1.3 Estimation basée sur l'imputation multiple

L'indice de valeur prédictive est établi en vue d'estimer $E(X)$ par imputation multiple appliquée à l'échantillon matriciel. Un estimateur fondé sur l'imputation multiple, \bar{y}_j de $E(Y)$ est approximé, sous l'hypothèse d'un nombre infini d'imputations, par

$$\bar{y}_j = \lim_{M \rightarrow \infty} M^{-1} \sum_{f=1}^M \bar{y}_j^f.$$

Dans cette expression, M est le nombre d'imputations, \bar{y}_j^f est la moyenne calculée d'après le j^{e} ensemble de données complet avec les valeurs imputées $X_{i,j}^f$, $i = 1, \dots, n_{\text{obs}}$, et les valeurs observées $X_{i,j}^f = X_{i,j}$ (qui ne varient pas d'un ensemble de données complet à l'autre), c'est-à-dire

$$\bar{y}_j^f = n_{\text{tot}}^{-1} \sum_{i=1}^{n_{\text{tot}}} X_{i,j}^f = n_{\text{tot}}^{-1} \left(\sum_{i=1}^{n_{\text{obs}}} X_{i,j}^f + \sum_{i=n_{\text{obs}}+1}^{n_{\text{tot}}} X_{i,j}^f \right).$$

Un estimateur de la variance de \bar{y}_j lorsque les imputations sont créées en utilisant X et Z , en se servant de la formule courante de la variance de Rubin (1987, section 3.1), est

$$V_{\text{IM}} = V_{\text{comp}} + V_{\text{imp}} \quad (1)$$

où le premier terme est une estimation de la variance que l'on obtiendrait avec les données complètes et le deuxième, une estimation de la variance entre les ensembles de données imputés. Dans le cas de grands échantillons, pour lesquels la variance de X peut être traitée comme étant connue, $V_{\text{comp}} = V(X)/n_{\text{tot}}$, et

provenant de la deuxième NHANES (NHANES II), puis le plan résultant et les méthodes d'imputation multiple sont évalués grâce à une étude par simulation portant sur les données de la NHANES III. La section 2 décrit la méthode de conception de questionnaires d'échantillonnage matriciel. La section 3 décrit le plan d'échantillonnage et les résultats de l'étude fondés sur la NHANES. Enfin, l'article se conclut par une discussion à la section 4.

2. Conception de questionnaires d'échantillonnage matriciel

La présente section décrit l'élaboration d'une méthode de conception de questionnaires d'échantillonnage matriciel, chacun contenant un sous-ensemble de questions destinées à être posées à un échantillon de répondants.

Lors de la conception d'un échantillon matriciel, il faut décider quelles questions seront considérées comme des questions communes qui seront incluses dans tous les questionnaires et quelles questions seront traitées comme des questions échantillonnées qui ne seront incluses que dans certains questionnaires. Habituellement, les questions communes sont sélectionnées en se basant sur un jugement de fonds et sur d'autres considérations quant à l'importance relative des questions. Les questions clés, pour lesquelles la précision de certains estimateurs doit être maximisée, devraient être traitées comme des questions communes, tandis que celles de moins grande importance peuvent être choisies comme questions échantillonnées. En outre, il est utile de sélectionner des questions communes qui sont prédictives d'un grand nombre de questions échantillonnées, afin que l'information sur les questions échantillonnées qui sont exclues d'un questionnaire puisse être récupérée d'après les questions communes combinées aux questions échantillonnées incluses dans le questionnaire. Enfin, le coût et le fardeau de réponse associés à une question sont aussi des éléments à prendre en considération, car il peut être avantageux de désigner des questions dont l'administration est coûteuse et (ou) représente un fardeau comme étant des questions échantillonnées. L'accent étant mis ici sur la façon de répartir les questions échantillonnées entre les questionnaires après que l'on ait choisi les questions communes, nous supposons que ces dernières ont déjà été sélectionnées. Cependant, nous verrons que la méthode de répartition des questions échantillonnées repose sur une mesure qui tient également compte de l'utilité des questions communes pour la prédiction des questions échantillonnées. Le pouvoir de prédire les valeurs des questions échantillonnées est estimé à l'aide d'un échantillon d'apprentissage.

Il faut aussi choisir un format pour l'organisation des questions échantillonnées. Afin de s'assurer que chaque paire possible de questions échantillonnées figure dans un

questionnaire, pour que l'estimation directe de toutes les associations bidirectionnelles entre variables soit possible, les questions échantillonnées sont réparties en blocs, et des questionnaires d'échantillonnage matriciel sont créés en assemblant deux ou plusieurs blocs de questions échantillonnées (Ragunathan et Grizzle 1995). La taille des blocs et leur nombre déterminent la longueur des questionnaires et leur nombre. Par exemple, dans l'étude portant sur la NHANES dont nous discuterons à la section 3, les questions échantillonnées sont réparties en quatre blocs, et chaque questionnaire contient deux blocs (ainsi que les questions communes), de sorte qu'il en existe six en tout (combinaisons de 2 parmi 4). Dans la méthode élaborée ici, les blocs sont approximativement de taille égale, et chaque question échantillonnée est attribuée à un seul bloc. L'utili-

sation de blocs de même longueur donne une réduction identique du fardeau de réponse pour tous les participants à l'étude. Elle produit aussi la même précision pour les questions de même type. Notons cependant que ces caractéristiques ne sont pas obligatoires pour tous les plans d'échantillonnage matriciel. Si une estimation de plus grande précision est souhaitée pour une question, celle-ci pourrait être incluse dans plus d'un questionnaire, ou être désignée comme une question commune devant figurer dans tous les questionnaires.

Un bon plan d'échantillonnage matriciel répartit les questions échantillonnées entre les blocs de façon que, pour toute question échantillonnée exclue d'un bloc, il existe des questions échantillonnées incluses dans le bloc qui, regroupées avec les questions communes, sont prédictives de la question exclue; cela facilite le recouvrement de l'information au sujet de la question exclue à l'étape de l'analyse des données. La discussion qui suit porte sur l'élaboration d'une méthode en vue d'atteindre cet objectif. Cette élaboration comporte deux volets. Premièrement, à la section 2.1, nous formulons un indice permettant de déterminer la mesure dans laquelle chaque question échantillonnée est prédite correctement par chaque autre question échantillonnée, l'utilité prédictive étant mesurée en tant que gain relatif de précision conditionnellement à l'inclusion des questions communes. Nous présentons aussi des méthodes d'estimation des valeurs de l'indice à partir d'un échantillon d'apprentissage. Deuxièmement, à la section 2.2, nous décrivons un algorithme pour l'affectation des questions échantillonnées aux blocs d'après l'indice de valeur prédictive.

2.1 Indice de valeur prédictive

2.1.1 Notation préliminaire pour les plans d'échantillonnage matriciel

Soit Y , une question échantillonnée pouvant être prédite, $X = (X_1, \dots, X_J)$, les questions communes et Z , une question échantillonnée utilisée pour prédire Y .

Le fait de ne poser qu'un sous-ensemble de questions de l'enquête à chaque répondant dans le cas d'un plan d'échantillonnage matriciel crée ce que l'on peut considérer comme des données manquantes au hasard, ou même manquant entièrement au hasard (Rubin 1976), puisqu'elles sont le résultat d'un mécanisme probabiliste commun fondé éventuellement sur des variables du plan de sondage. Le recours à l'imputation multiple (Rubin 1987), une approche polyvalente mise au point pour traiter les données présentant des valeurs manquantes, est tentant pour analyser les données provenant d'un échantillon matriciel, parce qu'après la création des imputations multiples, l'analyste peut appliquer les méthodes standard d'analyse des données complètes d'enquête par sondage. De surcroît, si l'échantillon matriciel a été conçu de telle façon que les questions posées à chaque répondant soient prédictives des questions qu'il n'a pas posées, alors, dans l'imputation multiple, il est possible d'utiliser les questions incluses pour recouvrer l'information au sujet de celles qui sont exclues. Nous nous concentrons sur l'imputation multiple, parce qu'elle est bien adaptée à cette situation : 1) l'application de méthodes multivariées complexes peut être exécutée une seule fois par l'organisme d'enquête qui connaît le mieux le plan de sondage; 2) l'imputation peut être implémentée au moyen de logiciels existants et 3) elle ne nécessite pas de nouvelles méthodes pour chacun des nombreux paramètres que ciblent la plupart des études. Néanmoins, d'autres méthodes d'estimation que l'imputation multiple, fondées sur un modèle ainsi que fondées sur un plan de sondage, peuvent être mises au point et appliquées aux données provenant de plans d'échantillonnage matriciel.

L'approche de l'échantillonnage matriciel a été appliquée et explorée dans diverses circonstances, comme l'évaluation des acquis scolaires (Sirotnik et Wellington 1977;

de ces sous-ensembles. Les différents sous-ensembles de questions (items) sont administrés à différents ensembles de répondants, afin que chaque question soit posée au moins à certains des répondants. Les plans d'enquête basés sur des questionnaires de ce type sont appelés plans à « questionnaire scindé » ou plans d'« échantillonnage matriciel », cette dernière expression reflétant l'idée que les répondants (lignes) et les questions (colonnes) sont les uns et les autres « échantillonnés » à partir d'une matrice conceptuelle de données sur la population complète. Dans de nombreux plans d'échantillonnage matriciel, certaines questions (appelées ici questions « communes ») sont posées à tous les répondants, tandis que d'autres (appelées ici questions « échantillonnées ») ne sont posées qu'à un sous-ensemble de répondants. Habituellement, les questions choisies comme questions communes sont soit particulièrement importantes, soit prédictives d'un grand nombre de questions échantillonnées.

Le présent article décrit la mise au point et l'évaluation d'une méthode de conception de questionnaires d'échantillonnage matriciel, chacun contenant un sous-ensemble de questions devant être posées à un échantillon aléatoire de répondants. La méthode peut être appliquée dans des conditions complexes, y compris les situations où il existe des enchevêtrements de questions. Les questionnaires sont conçus de telle façon que chacun comprenne des questions qui sont prédictives des questions exclues, de sorte que, lors des analyses subséquentes fondées sur l'imputation multiple, il soit possible de recouvrer l'information sur les questions exclues qui aurait été recueillie si l'on n'avait pas recouru à l'échantillonnage matriciel. La méthode suppose que l'on dispose d'un échantillon d'apprentissage. Ce dernier peut provenir de l'administration antérieure d'un questionnaire complet ou d'un échantillon pilote utilisé pour appuyer la conception du questionnaire. La méthode d'échantillonnage matriciel est évaluée dans le cadre d'une étude portant sur des données provenant de la National Health and Nutrition Examination Survey (NHANES), l'une des nombreuses enquêtes représentatives de la population nationale réalisée par le NCHS (<http://www.cdc.gov/nchs/nhanes.htm>). La NHANES, une enquête transversale qui a été répétée plusieurs fois au cours de diverses périodes, permet de recueillir une grande quantité de données auprès des répondants au moyen d'un questionnaire sur les membres des ménages, un examen médical dans un centre d'examen mobile et l'analyse en laboratoire de prélèvements biologiques. Il est intéressant d'étudier la faisabilité de plans d'échantillonnage matriciel pour des enquêtes telles que la NHANES, qui est caractérisée par des dépendances structurelles complexes entre les questions, ce que reflètent les nombreux enchevêtrements, ainsi que de multiples complications du questionnaire est appliquée à des données pilotes

Une évaluation des méthodes d'échantillonnage matriciel à l'aide de données provenant de la National Health and Nutrition Examination Survey

Neal Thomas, Trivellore E. Raghunathan, Nathaniel Schenker, Myron J. Katzoff et Clifford L. Johnson¹

Résumé

Les chercheurs et les responsables des politiques utilisent souvent des données provenant d'enquêtes par échantillonnage probabiliste représentatives de la population nationale. Le nombre de sujets couverts par ces enquêtes, et par conséquent la durée des entrevues, a généralement augmenté au fil des ans, ce qui a accru les coûts et le fardeau de réponse. Un remède éventuel à ce problème consiste à regrouper prudemment les questions d'un ensemble de sujets et à demander à chaque répondant de ne répondre qu'à l'un de ces sous-ensembles. Les plans de sondage de ce type sont appelés plans à « questionnaire schindé » ou plans d'« échantillonnage matriciel ». Le fait de ne poser qu'un sous-ensemble des questions d'une enquête à chaque répondant selon un plan d'échantillonnage matriciel crée ce que l'on peut considérer comme des données manquantes. Le recours à l'imputation multiple (Rubin 1987), une approche polyvalente mise au point pour traiter les données manquantes, est tentant pour analyser les données provenant d'un échantillon matriciel, parce qu'après la création des imputations multiples, l'analyste peut appliquer les méthodes standards d'analyse de données complètes provenant d'une enquête par sondage. Le présent article décrit l'élaboration et l'évaluation d'une méthode permettant de créer des questionnaires d'échantillonnage matriciel contenant chacun un sous-ensemble de questions devant être administrées à des répondants sélectionnés aléatoirement. La méthode peut être appliquée dans des conditions complexes, y compris les situations comportant des enchaînements de questions. Les questionnaires sont créés de telle façon que chacun comprenne des questions qui sont prédictives des questions exclues, afin qu'il soit possible, lors des analyses subséquentes fondées sur l'imputation multiple, de recouvrer une partie de l'information relative aux questions exclues qui aurait été recueillie si l'on n'avait pas recouru à l'échantillonnage matriciel. Ce dernier et les méthodes d'imputation multiple sont évalués aux moyens de données provenant de la National Health and Nutrition Examination Survey, l'une des nombreuses enquêtes par échantillonnage probabiliste représentatives de la population nationale réalisées par le National Center for Health Statistics des Centers for Disease Control and Prevention. L'étude démontre que l'approche peut être appliquée à une grande enquête nationale sur la santé à structure complexe et permet de faire des recommandations pratiques quant aux questions qu'il serait approprié d'inclure dans des plans d'échantillonnage matriciel lors de futures enquêtes.

Mots clés : Données manquantes; imputation multiple; fardeau de réponse; questionnaire schindé; enquête par sondage.

1. Introduction

Les données provenant d'enquêtes par sondage sont utilisées par les chercheurs et les responsables des politiques dans de nombreux domaines. Souvent, ces enquêtes mettent en jeu des échantillons probabilistes représentatifs de la population nationale et une collecte à grande échelle de données au moyen de questionnaires, et doivent concilier deux objectifs concurrents, c'est-à-dire d'une longueur et d'une complétude raisonnables tout en fournissant l'information pertinente. Le nombre de sujets couverts par ce genre d'enquêtes, et par conséquent la longueur des entrevues, ont généralement augmenté au fil des ans. L'accroissement résultant du fardeau de réponse pourrait être l'un des

facteurs qui contribuent à la diminution observée des taux de réponse. Cette baisse des taux de réponse risque de réduire la précision des estimations fondées sur les données d'enquête. Elle peut aussi accroître le biais, si les différences systématiques entre les non-répondants et les répondants ne sont pas prises en compte dans les analyses des données incomplètes. En outre, l'élargissement de la gamme de sujets couverts conjugué aux efforts en vue de maintenir les taux de réponse élevés ont accru le coût de la réalisation des enquêtes. Un moyen éventuel d'obtenir l'information nécessaire tout en limitant le fardeau de réponse consiste à regrouper prudemment les questions d'une enquête en sous-ensembles et à demander à chaque répondant de ne répondre qu'à l'un

1. Neal Thomas, DataMetrics Research, Inc., 61 Dream Lake Drive, Madison, CT 06443, E.-U. Courriel : snthomas99@yahoo.com; Trivellore E. Raghunathan, Department of Biostatistics and Institute for Social Research, University of Michigan, Ann Arbor, MI 48106, E.-U. Courriel : teraghnu@umich.edu; Nathaniel Schenker, Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, E.-U. Courriel : nschenker@cdc.gov; Myron J. Katzoff, Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, E.-U. Courriel : mkatzoff@cdc.gov; Clifford L. Johnson, Division of Health and Nutrition Examination Surveys, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, E.-U. Courriel : cjohnson@cdc.gov.

- BLS *Handbook of Methods* (2005). <http://stats.bls.gov/bls/descriptions.htm>.
- Consumer Price Indexes *Technical Manual* (2005). Office for National Statistics, London. http://www.statistics.gov.uk/downloads/theme_economy/CPI_Technical_Manual_2005.pdf.
- De Haan, J., Opperdoes, E. et Schut, C. (1999). Le choix des produits pour l'indice des prix à la consommation : Le seuil d'inclusion par opposition au sondage probabiliste. *Techniques d'enquête*, 25, 1, 33-45.
- Dalén, J. (1998). Studies on the comparability of consumer price indices. *Revue Internationale de Statistique*, 66, 1, 83-113.
- Diewert, E. (1997). "Commentary" [sur "Alternative Strategies for Aggregating Prices in the CPI" par M.D. Shapiro et D.W. Wilcox]. *Federal Reserve Bank of St. Louis Review*, 79, 3, 27-37.
- Diewert, E. (2004). Index number theory: Past progress and future challenges. Presented at the SSHRC Conference on Price Index Concepts and Measurement, Vancouver, Canada, à <http://www.econ.ubc.ca/diewert/concepts.pdf>.
- Dortman, A.H., Leaver, S.G. et Lent, J. (1999). Some observations on price index estimators. *Proceedings of the Federal Committee on Statistical Methodology Research Conference, Monday B Sessions*, 56-65.
- Reinsdorf, M., et Triplett, J.E. (2005). A review of reviews: Ninety years of professional thinking about the consumer price index. A parative, *Proceedings of the June 2004 NBER-CRUIW Conference on Price Indexes*, Vancouver.
- The Retail Prices Index *Technical Manual* (1998). (Ed. M. Baxter, The Stationary Office, London, à http://www.statistics.gov.uk/downloads/theme_economy/RPI_TECHNICAL_MANUAL.pdf.
- Richardson, D.H. (2000). Scanner indexes for the CPI. *Proceedings of the Conference on Scanner Data and Price Indexes*, NBER, Cambridge, <http://www.nber.org/books/>.
- Royal, R.M. (1976). Current advances in sampling theory: Implications for human observational studies. *American Journal of Epidemiology*, 104, 463-473.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model assisted Survey Sampling*. Springer, New York.
- Valliant, R., Dortman, A.H. et Royal, R.M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.

L'erreur quadratique moyenne comme étant un critère plus décisif, surtout sachant la bidirectionnalité des biais produits par *maxming*.

Tableau 14

Pourcentage d'ANE pour lequel l'indice de *Dutot* présente un biais positif pour un indice cible de Walsh pour deux scénarios d'échantillonnage

$ppt(\sqrt{q_y q_{y+1}})$	npe 2	npe 3	npe 4
1995 – 1996	75,0	86,2	75,9
1996 – 1997	60,7	72,4	65,5
1997 – 1998	65,5	75,9	78,6
1998 – 1999	72,4	75,9	70,4
1999 – 1900	89,7	72,4	75,9

Tableau 15

Pourcentage d'ANE pour lequel le biais et l'erreur quadratique moyenne de l'indice de *Dutot* pour un indice cible de Walsh sont plus faibles que l'indice cible de Walsh pour deux scénarios d'échantillonnage

a) Biais de ppt	b) EQM de ppt	plus faible	plus faible			
npe 2	npe 3	npe 4	npe 2 npe 3 npe 4			
1995 – 1996	82,1	93,1	86,2	32,1	58,6	41,4
1996 – 1997	89,2	96,6	100,0	35,7	37,9	27,6
1997 – 1998	89,7	86,2	100,0	41,4	24,1	64,3
1998 – 1999	89,7	82,8	92,6	41,4	37,9	40,7
1999 – 1900	89,7	96,6	41,4	34,5	31,0	37,9

Nous concluons que les bons effets de l'échantillonnage

maxming combinés à l'estimation de *Dutot* ne peuvent pas être expliqués par l'imitation approximative de l'échantillonnage *ppt*. Les comportements sont différents, et, dans l'ensemble, *maxming* semble être un peu meilleur que *ppt*.

Nous ne voyons aucune autre raison expliquant pourquoi l'indice d'échantillon de *Dutot* devrait cibler l'indice de population de Walsh lorsque les produits les mieux vendus sont échantillonnés systématiquement, à part le mécanisme de « force brute » : dans la mesure où l'indice de Walsh peut être représenté par un petit échantillon d'articles, il est mieux représenté par ceux pour lesquels les quantités sont systématiquement les plus grandes, et ces articles sont ceux que fournit le scénario d'échantillonnage *maxming*.

Bibliographie

Balk, B. (1999). On the use of unit values as consumer price subindices. *Proceedings of the Fourth Meeting of the International Working Group on Price Indices*, BLS, Washington, D.C.

Balk, B. (2003). Price indexes for elementary aggregates: The sampling approach. *Proceedings of the Seventh Meeting of the International Working Group on Price Indices* (Ottawa Group), Paris.

où $E_n()$ signifie l'espérance par rapport au plan d'échantillonnage et L_p est un indicateur aléatoire prenant la valeur de 1 ou 0, si i' est dans l'échantillon ou non. Nous obtenons une expression similaire pour le dénominateur. Le ratio de ces deux valeurs espérées est l'indice de Walsh. Par conséquent, à part le (léger) biais de ratio habituel, qui, comme nous pouvons le montrer, est généralement positif, l'indice de *Dutot* cible en effet l'indice de Walsh sous le scénario d'échantillonnage *ppt*.

Nous devons nous demander si les deux modes d'échantillonnage ont effectivement tendance à présenter un chevauchement appréciable en ce qui concerne les articles sélectionnés. Pour chaque passage-machin, pour chaque *npe* *l*, *ANE* *c*, trois articles ont été sélectionnés par échantillonnage *maxming* ou par échantillonnage *ppts* ($\sqrt{q_y q_{y+1}}$) des articles compris dans *lc*. Le tableau 13 donne le pourcentage de fois (sur 500 passages) que des articles différents ont été sélectionnés dans l'échantillon, pour certains cas représentatifs sélectionnés arbitrairement. Nous concluons, pas entièrement sans nous étonner, que a) l'échantillonnage *ppt* produit une plus grande dispersion des articles sélectionnés, b) les articles sélectionnés par échantillonnage *maxming* constituent un sous-ensemble de ceux sélectionnés par échantillonnage *ppt*, c) il existe une certaine corrélation des « articles dominants », c'est-à-dire de ceux qui ont le plus tendance à être sélectionnés par l'une ou l'autre méthode. Brevement, les échantillonnages *maxming* et *ppt* ($\sqrt{q_y q_{y+1}}$) semblent être reliés, mais lâchement.

Afin de mieux comprendre la relation entre les deux méthodes d'échantillonnage, nous avons estimé le biais et l'erreur quadratique moyenne, par rapport à l'indice de population de Walsh, de l'indice de *Dutot* pour chaque ANE, pour l'échantillonnage *maxming* ainsi que *ppt* ($\sqrt{q_y q_{y+1}}$). Les estimations du biais et de l'EQM étaient fondées sur 500 passages pour chaque méthode d'échantillonnage. Les statistiques sommaires ont été calculées sur l'ensemble des ANE pour chaque paire d'années et chaque *npe*. Le tableau 14 donne le pourcentage d'ANE pour lequel les indices élémentaires de *Dutot* présentent un biais positif pour chaque mode d'échantillonnage. Comme prévu, l'échantillonnage *ppt* a tendance à produire un biais positif, nous constatons que l'échantillonnage *maxming* donne des résultats tout aussi biaisés positivement que négativement.

Le tableau 15 (a) donne le pourcentage d'ANE pour lequel le biais absolu dû à l'utilisation de *maxming* est plus important que celui du *ppt* ($\sqrt{q_y q_{y+1}}$). À cet égard, l'échantillonnage *ppt* est meilleur. Cependant le tableau 15(b) donne le pourcentage d'ANE pour lequel *maxming* produit une plus grande erreur quadratique moyenne et, ici, *maxming* donne de meilleurs résultats pour toutes les combinaisons période/*npe*, sauf deux. Nous considérons

Törnqvist

$$T = \prod_{i=1}^I \left(\frac{d_{i,y}^i}{d_{i,y+1}^i} \right)^{s_{i,y,y+1}^i}$$

où

$$s_{i,y,y+1}^i = \frac{1}{2} \left(\frac{\sum_{i=1}^I d_{i,y}^i b_{i,y}^i}{d_{i,y+1}^i d_{i,y+1}^i} + \frac{\sum_{i=1}^I d_{i,y+1}^i b_{i,y+1}^i}{d_{i,y}^i d_{i,y}^i} \right)$$

Moyenne géométrique

$$G = \prod_{i=1}^I \left(\frac{d_{i,y+1}^i}{d_{i,y}^i} \right)$$

où

$$w_i = s_{i,y}^i = \left(\frac{\sum_{i=1}^I d_{i,y}^i b_{i,y}^i}{d_{i,y}^i d_{i,y}^i} \right)$$

ou

$$w_i = 1/N$$

Valeur unitaire

$$U = \frac{\sum_{i=1}^I q_{i,y+1}^i d_{i,y+1}^i / \sum_{i=1}^I q_{i,y}^i d_{i,y}^i}{\sum_{i=1}^I q_{i,y+1}^i d_{i,y+1}^i / \sum_{i=1}^I q_{i,y}^i d_{i,y}^i}$$

Dutot

$$RM = \frac{\sum_{i=1}^I d_{i,y+1}^i / N}{\sum_{i=1}^I d_{i,y}^i / N}$$

(« rapport des moyennes arithmétiques »)

$$MR = \frac{\sum_{i=1}^I d_{i,y+1}^i / d_{i,y}^i}{N}$$

Moyenne des ratios

Annexe B

Exemple illustrant l'importance du niveau le plus faible d'agrégation

Nous présentons ici un exemple simple en vue d'illustrer l'importance de la méthode utilisée pour construire les indices élémentaires. Nous comparons les indices de population de Walsh aux indices résultant de l'agrégation des indices élémentaires de Walsh selon une formule de Laspeyres. La raison pour laquelle nous nous concentrons sur l'indice de Walsh est donnée à l'annexe C. L'indice de Walsh « pur » est

$$E^w = \left(\sum_{i \in I} d_{i,y+1}^i \right) \left(\sum_{i \in I} d_{i,y}^i \right)^{-1} = \frac{\sum_{i \in I} \sqrt{d_{i,y}^i d_{i,y+1}^i} b_{i,y}^i}{\sum_{i \in I} \sqrt{d_{i,y}^i d_{i,y+1}^i} d_{i,y}^i}$$

avons

Pourquoi la combinaison *maxming/Dutot* donne-t-elle d'aussi bons résultats, paraissant donner lieu à une absence de biais pour les indices superlatifs?

Un examinateur nous a fait remarquer que l'échantillonnage *maxming* ressemble considérablement à l'échantillonnage *ppt* avec taille variable $\sqrt{q_{i,y}^i q_{i,y+1}^i}$; pour l'échantillonnage *ppt* sans biais pour un indice cible de Walsh et, par conséquent, indirectement, pour tout autre indice superlatif.

En effet, pour l'espérance du numérateur de l'indice de *Dutot*, sous le scénario d'échantillonnage probabiliste, nous

Annexe C
La combinaison *maxming/Dutot*

indice.

Les résultats sont donnés au tableau 9. Nous observons un écart perceptible entre l'indice de population réelle de Walsh et l'agrégat selon Laspeyres des indices élémentaires de Walsh, celui-ci ayant tendance à être un peu plus élevé. Cependant, ces différences sont du même ordre que celles comparativement à l'écart entre la *moyenne géométrique* ou l'indice de Laspeyres et les indices superlatifs. Ce genre de résultat confirme qu'une procédure valable au niveau le plus faible est un élément essentiel de la construction d'un

Les résultats sont donnés au tableau 9. Nous observons un écart perceptible entre l'indice de population réelle de Walsh et l'agrégat selon Laspeyres des indices élémentaires de Walsh, celui-ci ayant tendance à être un peu plus élevé.

où $w_{i,y,y+1}^h$ est le h^e indice élémentaire de Walsh et

$$W = \frac{\sum_{i=1}^I \sqrt{q_{i,y}^i q_{i,y+1}^i} d_{i,y}^i}{\sum_{i=1}^I \sqrt{q_{i,y}^i q_{i,y+1}^i} d_{i,y+1}^i} = \sum_{i=1}^I \tilde{s}_{i,y,y+1}^h w_{i,y,y+1}^h$$

quadratique moyenne relative de l'approche du Royaume-Uni deviendrait de plus en plus faible.

En pratique, évidemment, les quantités de la période 2 ne sont pas disponibles au moment de la sélection de l'échantillon (la période 1) et, dans le cadre de notre étude de suivi, nous donnons une certaine idée de la dégradation partielle qui résulte de l'utilisation des quantités antérieures : elle n'est pas suffisamment importante pour empêcher de conclure que l'approche du Royaume-Uni donne de meilleurs résultats. De surcroît, le jugement de l'économiste de terrain quant au meilleur vendeur pourrait être fondé sur des données plus récentes que celles d'il y a un an. Donc, l'effet réel pourrait être compris entre ceux des versions décalées et non décalées de *maxming* que nous avons utilisées. Toutefois, en pratique, les économistes de terrain des États-Unis pourraient échantillonner fréquemment des articles dans les points de vente d'après une estimation de la part des dépenses qui est réellement une moyenne lissée des parts des dépenses de la période de base et de la période récente. Cela pourrait atténuer le biais que nous avons observé dans nos simulations, où seules les dépenses de la période de base ont été utilisées pour l'échantillonnage dans les magasins.

Tableau 11

Biais, écart-type et racine de l'erreur quadratique moyenne (tous multipliés par 1 000), dans l'estimation de l'indice de population de Walsh pour l'ensemble des céréales, chaîne 8, fondée sur trois approches d'échantillonnage/estimation des indices élémentaires*

	a) Biais		
	1995 – 1996	1996 – 1997	1997 – 1998
<i>Dutot/maxming</i>	29	15	-13
<i>Dutot/maxming, q antérieures</i>	-	46	32
Moyenne géométrique/pplar	78	62	66

	b) Écart-type		
	1995 – 1996	1996 – 1997	1997 – 1998
<i>Dutot/maxming</i>	16	13	11
<i>Dutot/maxming, q antérieures</i>	-	14	12
Moyenne géométrique/pplar	22	18	17

	c) Racine de l'erreur quadratique moyenne		
	1995 – 1996	1996 – 1997	1997 – 1998
<i>Dutot/maxming</i>	33	20	17
<i>Dutot/maxming, q antérieures</i>	-	48	34
Moyenne géométrique/pplar	80	65	68

* Au niveau de l'AN/E/article représentatif. Pour obtenir des estimations de l'indice global, nous avons agréé les estimations des indices élémentaires en utilisant les dépenses de population connues.

Tableau 12

Ratios de la REQM du R.-U. à la REQM des E.-U., chaîne 8, indices cibles de Walsh : *maxming* en utilisant les valeurs décalées de *q* et *Dutot* versus *pplar*(dépenses) et moyenne géométrique

Description	1996 – 1997	1997 – 1998	1998 – 1999	1999 – 2000
Tous les articles	0,748	0,498	0,993	0,567
Catégories/Grands groupes	1,539	0,495	1,280	0,765
1 – Chaudes	0,563	0,676	0,941	0,797
2 – Sucrées	0,409	0,223	0,463	0,852
3 – Fruitées	0,915	0,560	1,164	0,359
4 – Ordinaires	0,748	0,607	0,660	0,657
Chaudes – 11	1,695	0,599	1,333	0,843
Sucrées – 21	0,757	0,593	1,136	0,924
Sucrées – 22	0,370	0,776	0,751	0,671
Sucrées – 23	0,479	0,785	0,796	0,508
Fruitées – 31	0,570	0,443	0,678	1,008
Fruitées – 32	0,526	0,350	0,277	0,674
Ordinaires – 41	1,167	0,509	1,395	0,397
Ordinaires – 42	0,623	0,411	0,918	0,624
Ordinaires – 43	0,919	1,171	0,668	0,560

Néanmoins, en ce qui a trait aux trois mesures d'exactitude (biais, écart-type et racine de l'erreur quadratique moyenne), la combinaison *maxming/Dutot* du Royaume-Uni continue de donner de meilleurs résultats que l'approche des États-Unis représentant l'échantillonnage probabiliste.

Pour les catégories plus fines, le tableau 12 donne les ratios des erreurs quadratiques moyennes obtenues sous la méthode du Royaume-Uni avec les valeurs décalées de q à celles obtenues sous la méthode des États-Unis. Bien qu'ils soient généralement plus élevés que ceux du tableau 8, ils donnent encore à penser que l'approche de l'échantillonnage par choix raisonné du Royaume-Uni est meilleure.

6. Discussion

Nous avons présenté une comparaison de deux approches fondamentalement différentes de l'échantillonnage et de l'inférence pour l'établissement d'un indice des prix à la consommation. La conclusion inévitable est que, dans la population que nous avons étudiée, l'approche « R-U », qui comporte une stratification plus stricte et, par-dessus tout, un échantillonnage au jugé dans les strates plus restrictif que l'échantillonnage probabiliste de l'approche « E-U », produit de meilleurs estimations d'un indice superlatif cible.

Nous montrons qu'il en est ainsi, quel que soit l'estimateur de l'indice de prix de faible niveau (*Dutot*, ou *moyenne géométrique*, ou la moyenne des ratios) employé, bien que le *Dutot* (rapport des moyennes) donne les meilleurs résultats.

L'approche du Royaume-Uni est supérieure pour deux raisons : 1) son échantillonnage plus strict, limité aux articles sélectionnés (par exemple, voir le tableau 13 décrit à l'annexe C), même, sans surprise, à une variance plus faible, constatation qui avait déjà été faite par de Haan et coll. (1999) et 2) les indices d'échantillon de *Dutot* ciblent les indices superlatifs sous échantillonnage du marché dominant, ce qui nous a surpris et a suscité l'étude décrite à la section 5. Par ailleurs, l'approche des États-Unis a donné un estimateur de l'indice pouvant être décrit comme étant sans biais, mais il était sans biais pour le « mauvais » indice de population basé sur la *moyenne géométrique* pondérée par les dépenses à la première période. Donc, il avait tendance à être considérablement plus élevé que l'indice superlatif cible, qu'il s'agisse de celui de Fisher, de Walsh ou de Törnqvist).

Si nous permetions aux tailles d'échantillon d'augmenter, nous pourrions nous attendre à ce que les variances des approches américaine et britannique diminuent l'une et l'autre, mais la variance du Royaume-Uni demeurerait plus faible. Le biais de l'estimateur des États-Unis pour l'indice superlatif cible ne serait pas affecté par l'accroissement de la taille d'échantillon, de sorte que l'erreur

Le fait que l'indice d'échantillon de *Dutot* puisse cibler l'indice de population de Walsh (et donc, indirectement, tout indice superlatif) lorsque les vendeurs les plus importants sont systématiquement échantillonnés est, selon nous, le résultat d'un mécanisme très simple, de « force brute » : dans la mesure où l'indice de Walsh peut être représenté par un petit échantillon d'articles, il est représenté le mieux par ceux pour lesquels les quantités sont régulièrement les plus grandes, et ce sont ces articles que le scénario d'échantillonnage *maxming* fournit presque tous les jours. À l'annexe C, nous discutons d'une autre explication des bonnes propriétés de la combinaison *maxming/Dutot*.

La moyenne des erreurs quadratiques moyennes a également été calculée pour la combinaison *maxming/Dutot* en se fondant sur les valeurs *antérieures* de q , c'est-à-dire q_{j-1}^i , q_{j-1}^v . Les résultats sont présentés à la dernière ligne du tableau 10. Nous observons un affaiblissement attendu comparativement à la combinaison *maxming/Dutot* mise à jour, mais la comparaison des résultats à ceux d'autres options demeure favorable. Nous étudions cet aspect plus en profondeur à la sous-section 5.2.

5.2 Effet des quantités décalées sur l'échantillonnage *maxming*

Après de placer les résultats de la section 4 dans leur contexte, nous devons déterminer quel est l'effet de l'utilisation de valeurs décalées de q dans *maxming*. La raison en est simple : si, à première vue, l'utilisation des quantités des périodes de base et courante semble le moyen évident de refléter la notion d'articles persistants utilisée au Royaume-Uni, cela implique néanmoins l'utilisation d'information (les quantités de la période courante) qui n'a pas été utilisée pour simuler l'échantillonnage utilisé aux États-Unis, ce qui pourrait donner un avantage injuste à la méthode du Royaume-Uni.

Par conséquent, nous comparons l'approche des États-Unis, c'est-à-dire *ppdvar* (avec taille variable basée sur les dépenses de la période) à la *moyenne géométrique* au niveau élémentaire à l'approche du Royaume-Uni représentée par *maxming* – *Dutot*, mais en fondant ici *maxming* sur les quantités q_{j-1}^i et q_{j-1}^v . Nous avons réduit légèrement les ensembles de données pour être certains d'obtenir des données concordantes pour les trois années consécutives. L'aggrégation pour produire des indices de niveau plus élevé a été faite d'après les dépenses réelles de population pour les États-Unis, ainsi que le Royaume-Uni.

Le tableau 11 donne les résultats pour les indices calculés pour l'ensemble des céréales pour la chaîne 8, en comparant les biais, les écarts-types et les racines de l'erreur quadratique moyenne par rapport à l'indice de population de Walsh. Comme prévu, les résultats ne sont pas aussi bons que ceux obtenus en utilisant les valeurs courantes de q .

Pour chaque mode d'échantillonnage, dans chaque combinaison *upc/ANF*, nous avons tiré 500 échantillons. Nous avons calculé l'erreur quadratique moyenne des estimations par rapport à un indice cible de Walsh au niveau de l'ANF. Les moyennes des *egm* sur l'ensemble des ANF ont été calculées pour chaque mode d'échantillonnage/estimation, dans chaque *upc*.

Le tableau 10 donne le ratio de ces moyennes par rapport à l'*egm* moyenne pour la combinaison *maxming/Dutot*. Pour chaque estimateur, pour chaque *upc*, à une exception près (*upc* 3, 1999-2000), *maxming* donne l'*egm* la plus faible, souvent avec une marge appréciable. L'échantillonnage *ppt* sans remise est la deuxième des solutions les meilleures. Si nous maintenons la méthode d'échantillonnage fixe (en

Tableau 9
Indices de population 1995-1996, chaîne 8

Description	Laspeyres	Moyenne géométrique*	Fisher	Walsh	Laspeyres de l'indice élémentaire de Walsh
Tous les articles	1,129	1,091	1,028	1,030	1,040
Catégories/Grands groupes					
1 - Chaudes	1,161	1,115	1,080	1,082	1,084
2 - Sucrées	1,129	1,088	1,007	1,012	1,025
3 - Fruitées	1,084	1,054	0,997	1,005	1,015
4 - Ordinaires	1,135	1,101	1,046	1,042	1,050
Strates d'articles/Sections					
Chaudes - 11	1,157	1,117	1,088	1,089	1,090
Chaudes - 12	1,164	1,113	1,072	1,075	1,079
Sucrées - 21	1,086	1,045	0,962	0,970	0,992
Sucrées - 22	1,187	1,142	1,055	1,056	1,058
Sucrées - 23	1,117	1,091	1,034	1,039	1,043
Fruitées - 31	1,003	0,992	0,949	0,965	0,966
Fruitées - 32	1,228	1,172	1,100	1,091	1,102
Ordinaires - 41	1,212	1,161	1,091	1,080	1,090
Ordinaires - 42	1,048	1,030	0,997	0,997	0,998
Ordinaires - 43	1,136	1,107	1,048	1,046	1,056

* Pondérée par les dépenses à la période de base.

Tableau 10
Moyenne standardisée de l'erreur quadratique moyenne relative sur l'ensemble des ANF, populations réduites, chaîne 8

	<i>upc</i> 2				<i>upc</i> 3				<i>upc</i> 4			
Estimateur/méthode d'échantillonnage	96-97	97-98	98-99	99-00	96-97	97-98	98-99	99-00	96-97	97-98	98-99	99-00
<i>Dutot/pptst</i>	1,73	1,70	1,68	1,91	1,23	1,82	1,35	2,24	1,22	1,06	1,12	0,93
<i>Dutot/pptst</i>	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
<i>Dutot/pptst</i>	2,13	2,10	1,91	2,14	1,42	2,10	1,46	2,67	1,45	1,23	1,36	1,07
Moyenne géométrique/ <i>maxming</i>	1,20	1,16	1,16	1,06	1,06	1,14	1,08	1,05	1,10	1,11	1,12	0,96
Moyenne géométrique/ <i>pptst</i>	2,08	1,88	1,98	2,27	1,33	1,94	1,47	2,59	1,33	1,09	1,28	0,97
Moyenne géométrique/ <i>pptst</i> (E.-U.)	2,49	2,29	2,18	2,53	1,58	2,23	1,58	3,09	1,59	1,30	1,52	1,12
<i>RM/maxming</i>	1,42	1,32	1,31	1,14	1,24	1,03	1,30	1,05	1,11	1,20	1,21	1,07
<i>RM/pptst</i>	2,81	2,35	2,49	2,85	1,70	2,31	1,77	3,43	1,57	1,30	1,42	1,17
<i>RM/pptst</i>	3,23	2,77	2,66	3,08	2,03	2,58	1,87	3,96	1,83	1,49	1,66	1,30
<i>Dutot/maxming, g antérieures</i>	1,12	1,19	1,19	1,19	1,56	1,42	1,69	1,51	1,20	1,02	0,85	1,48

5. Étude de suivi

5.1 Méthodes d'échantillonnage et estimateurs au niveau élémentaire

Pour explorer ces questions, nous avons exécuté d'autres études par simulation. Nous nous sommes servis des mêmes données sur les céréales que celles utilisées pour l'étude principale (mois de février successifs), mais nous avons limité aux magasins indépendants, *chaîne* 8. Nous avons procédé ainsi pour que l'étude soit plus facile à gérer, mais aussi parce que, pour les autres chaînes, les estimateurs des indices élémentaires utilisés au Royaume-Uni étaient plus compliqués que le simple indice de *Dwot*. En outre, il est raisonnable de s'attendre à ce que le comportement des prix soit le plus hétérogène dans cette chaîne, de sorte que les différences intrinsèques seront plus évidentes. La chaîne 8 est la plus grande des chaînes étudiées, comprenant chaque année environ 30 % de l'ensemble de la population, soit environ 6 000 enregistrés.

La structure de base est restée la même : 3 *upc*, 4 grands groupes/catégories de dépenses (chaudes, sucrées, fruitées et ordinaires), 10 sections/strates d'articles, et 29 articles représentatifs/*ANF*. Pour chaque *ANF/article représentatif*, 3 points de vente (un article par point de vente) ont été sélectionnés, par opposition à 10 dans le cas de l'étude principale. Pour étudier l'approche *maximiq* basée sur des périodes antérieures, nous avons réduit les cinq ensembles de données originaux, contenant chacun les données sur les prix et les quantités pour une paire d'années (1995–1996, 1996–1997, etc.) afin de n'inclure que les articles permettant un « rétro-appariement » : c'est-à-dire l'appariement sur trois années pour comparer les prix des articles dans les points de vente pour 1995/1996/1997, 1996/1997/1998, etc. Environ 90 % des enregistrés de la chaîne 8 ont permis un rétro-appariement (en ce qui concerne les résultats qui suivent, il vaut probablement la peine de souligner que la réduction de l'échantillon pourrait influencer de façon disproportionnée le *maximiq*). Nous déplaçons notre attention de l'indice de Fisher vers l'indice superlatif de Walsh, grâce à une suggestion astucieuse d'un examinateur, dont nous discutons à l'annexe C.

Nous avons utilisé trois estimateurs pour les indices élémentaires : le ratio des moyennes arithmétiques (RM) (le *Dwot*), la *moyenne géométrique* non pondérée (aussi appelée indice de Jevons) et la moyenne des ratios (*MRR*). Dans l'échantillonnage *ppt* des points de vente, puis dans l'échantillonnage des articles dans les points de vente, nous avons supposé que la variable de taille (dépenses) était connue (au lieu d'être estimée, comme dans l'étude principale). Outre l'échantillonnage *ppt* avec remise (comme dans l'approche américaine), et *maximiq*, nous avons également étudié l'échantillonnage *ppt* sans remise, parce que nous soupçonnions qu'il serait moins variable que la version avec remise.

Les approches du Royaume-Uni et des États-Unis diffèrent à quatre égards : 1) la structure de stratification, en particulier l'utilisation au Royaume-Uni de différentes strates de magasins et, dans une certaine mesure, de l'échantillonnage centralisé, 2) la structure d'agrégation et de pondération, 3) le mode d'échantillonnage à divers degrés et 4) la formule des agrégats élémentaires. Il est donc difficile de déterminer dans quelle mesure chaque aspect contribue au mérite relatif des méthodes américaine et britannique de construction de l'indice. En particulier, comme nous l'avons mentionné à la dernière section, la raison pour laquelle l'estimateur de l'indice du Royaume-Uni a tendance à cibler les indices superlatifs, surtout au niveau plus élevé d'agrégation, demeure un peu mystérieuse.

Dans l'étude de suivi, nous nous concentrons sur le niveau le plus faible de construction de l'indice, c'est-à-dire sur (3). Le niveau magasin-article représentatif (*ANF*) d'échantillonnage et sur (4), les formules des indices élémentaires. Nous comparons les avantages relatifs des diverses options, en prenant comme cibles les indices élémentaires à options, en prenant comme cibles les indices élémentaires à points de vente (un article par point de vente) ont été sélectionnés, par opposition à 10 dans le cas de l'étude principale. Pour étudier l'approche *maximiq* basée sur des périodes antérieures, nous avons réduit les cinq ensembles de données originaux, contenant chacun les données sur les prix et les quantités pour une paire d'années (1995–1996, 1996–1997, etc.) afin de n'inclure que les articles permettant un « rétro-appariement » : c'est-à-dire l'appariement sur trois années pour comparer les prix des articles dans les points de vente pour 1995/1996/1997, 1996/1997/1998, etc. Environ 90 % des enregistrés de la chaîne 8 ont permis un rétro-appariement (en ce qui concerne les résultats qui suivent, il vaut probablement la peine de souligner que la réduction de l'échantillon pourrait influencer de façon disproportionnée le *maximiq*). Nous déplaçons notre attention de l'indice de Fisher vers l'indice superlatif de Walsh, grâce à une suggestion astucieuse d'un examinateur, dont nous discutons à l'annexe C.

Nous avons utilisé trois estimateurs pour les indices élémentaires : le ratio des moyennes arithmétiques (RM) (le *Dwot*), la *moyenne géométrique* non pondérée (aussi appelée indice de Jevons) et la moyenne des ratios (*MRR*). Dans l'échantillonnage *ppt* des points de vente, puis dans l'échantillonnage des articles dans les points de vente, nous avons supposé que la variable de taille (dépenses) était connue (au lieu d'être estimée, comme dans l'étude principale). Outre l'échantillonnage *ppt* avec remise (comme dans l'approche américaine), et *maximiq*, nous avons également étudié l'échantillonnage *ppt* sans remise, parce que nous soupçonnions qu'il serait moins variable que la version avec remise.

Donc, une source vraisemblablement importante de la différence entre les résultats donnés par les méthodes américaine et britannique tient à l'estimation sur échantillon des indices élémentaires de population. Mais cela laisse ouverte la question de savoir si les écarts sont dus à des différences entre les méthodes d'échantillonnage ou entre les formules utilisées pour l'estimation, ou les deux. Donc, nous cherchons à déterminer 1) comment l'échantillonnage au jugé (ici, échantillonnage avec seuil d'inclusion basé sur *maximiq*) se compare à l'échantillonnage probabiliste représenté par *ppdw*, en maintenant fixe l'estimateur des indices élémentaires, et 2) comment les estimateurs des indices élémentaires se comportent lorsque nous maintenons fixe la méthode d'échantillonnage. Il sera également intéressant de déterminer ce qui se passe lorsque l'échantillonnage *maximiq* est fondé sur des données provenant de la période de base et de la période *précédente*, plutôt que de la période de base et la période courante.

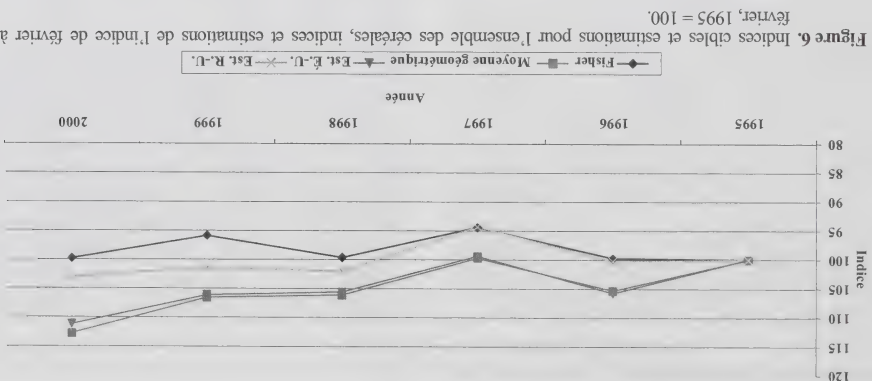


Figure 7. Différences entre les estimations britanniques et les indices de population de Fisher, indices et estimations de l'indice de février à février, 1995 = 100.

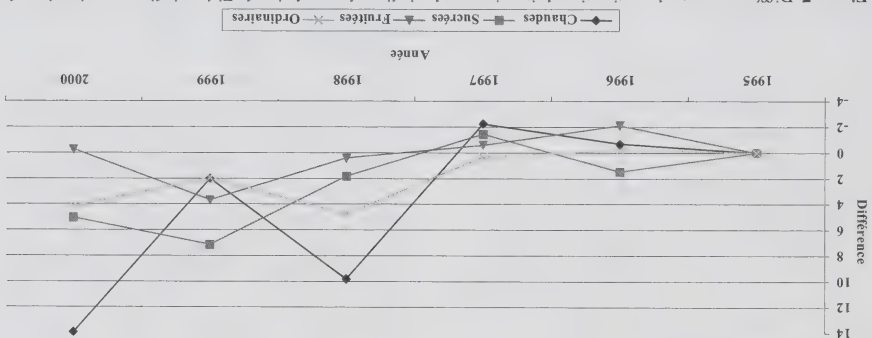


Tableau 8
Ratios de la REQM du R.-U. à la REQM des E.-U.

Description	1995-1996	1996-1997	1997-1998	1998-1999	1999-2000
Tous les articles	0,196	0,192	0,419	0,548	0,288
Catégories/Grands groupes					
1 - Chaudes	0,713	0,517	0,483	1,437	0,589
2 - Sucrées	0,286	0,336	0,314	0,522	0,282
3 - Fruitées	0,595	0,508	0,308	0,501	0,405
4 - Ordinaires	0,310	0,297	0,777	0,319	0,404
Secteurs d'articles/Sections					
Chaudes - 11	0,923	1,066	0,682	0,529	0,508
Chaudes - 12	0,920	0,850	1,169	1,860	0,842
Sucrées - 21	0,702	0,392	0,421	0,595	0,330
Sucrées - 22	1,092	0,426	0,380	0,341	0,365
Sucrées - 23	0,650	0,455	0,448	0,925	0,851
Fruitées - 31	0,778	1,059	0,637	0,581	0,618
Fruitées - 32	0,683	0,809	0,314	0,457	0,356
Ordinaires - 41	0,709	0,623	0,494	0,567	0,317
Ordinaires - 42	0,642	0,511	1,117	1,092	1,005
Ordinaires - 43	0,678	0,839	0,641	0,815	0,701

étionnant, moins de variance. Néanmoins, cela est également dû, en partie, à une tendance étonnante des estimateurs britanniques à cibler les indices de Fisher correspondants, ce qui réduit le biais. Puisque les estimateurs britanniques ne ressemblent pas formellement à l'indice de Fisher, les raisons de leur tendance à approximer cet indice méritent d'être étudiées plus en profondeur. Nous nous penchons sur cette question à la section suivante.

Tableau 6
Indices 1995-1996 cibles possibles

Description	Moyenne géométrique	Törnqvist	Fisher	Laspeyres
Tous les articles	1,053	1,002	0,997	1,079
Catégories/Grands groupes				
1 - Chaudes	1,058	1,052	1,052	1,078
2 - Sucrées	1,042	0,964	0,956	1,072
3 - Fritées	1,044	1,007	1,007	1,067
4 - Ordinaires	1,069	1,027	1,027	1,092
Strates d'articles/Sections				
Chaudes - 11	1,043	1,044	1,044	1,057
Chaudes - 12	1,073	1,059	1,058	1,097
Sucrées - 21	1,003	0,917	0,910	1,034
Sucrées - 22	1,063	0,982	0,972	1,093
Sucrées - 23	1,093	1,052	1,054	1,119
Fritées - 31	0,977	0,955	0,950	0,985
Fritées - 32	1,165	1,110	1,116	1,204
Ordinaires - 41	1,067	1,021	1,021	1,094
Ordinaires - 42	1,030	0,996	0,996	1,050
Ordinaires - 43	1,104	1,063	1,062	1,125

Tableau 7
Résultats des simulations pour les indices 1995-1996

Description	Indice cible	Moyenne	Ecart-type	REQM	Moyenne	Ecart-type	REQM	R-U.
Tous les articles	0,997	1,057	0,016	0,062	1,002	0,011	0,012	
Catégories/Grands groupes								
1 - Chaudes	1,052	1,059	0,031	0,032	1,045	0,022	0,023	
2 - Sucrées	0,956	1,046	0,030	0,095	0,971	0,023	0,027	
3 - Fritées	1,007	1,053	0,035	0,058	0,986	0,027	0,034	
4 - Ordinaires	1,027	1,072	0,025	0,051	1,025	0,016	0,016	
Strates d'articles/Sections								
Chaudes - 11	1,044	1,045	0,035	0,035	1,064	0,025	0,032	
Chaudes - 12	1,058	1,072	0,049	0,051	1,027	0,035	0,047	
Sucrées - 21	0,910	1,004	0,050	0,106	0,850	0,045	0,074	
Sucrées - 22	0,972	1,070	0,051	0,111	1,089	0,030	0,121	
Sucrées - 23	1,054	1,095	0,044	0,060	1,026	0,027	0,039	
Fritées - 31	0,950	0,979	0,020	0,035	0,932	0,020	0,027	
Fritées - 32	1,116	1,178	0,084	0,104	1,077	0,059	0,071	
Ordinaires - 41	1,021	1,069	0,050	0,070	1,060	0,030	0,049	
Ordinaires - 42	0,996	1,033	0,035	0,051	0,987	0,031	0,032	
Ordinaires - 43	1,062	1,107	0,042	0,061	1,028	0,023	0,041	

4. Résultats de l'étude principale

Le tableau 6 donne les indices comparant 1995 à 1996 pour la population (1) dans son ensemble (les trois domaines combinés), (2) ventilée en classes/grands groupes et (3) ventilée encore davantage en strates d'articles/sections. Quatre indices, qui pourraient être considérés comme les cibles de l'estimation, sont donnés. À cet égard, rappelons la discussion sur les cibles qui figurent à la section 1.

Le tableau 7 donne les moyennes, variances et erreurs quadratiques moyennes correspondantes pour les estimations américaines et britanniques, où l'erreur quadratique moyenne est calculée par rapport aux indices de Fisher. Nous faisons les constatations suivantes :

- (1) Pour l'ensemble des articles, les catégories et les strates d'articles, les estimations américaines semblent s'approcher de la *moyenne géométrique* G. Cela confirme ce que d'autres travaux nous avaient fait soupçonner (Dorfman et coll., 1999), à savoir que le niveau le plus faible d'aggrégation est dominant (nous avons utilisé une formule de Laspeyres pour les niveaux d'aggrégation plus élevés). Le fait que G se situe entre l'indice de Laspeyres et l'indice cible superlatif donne certaines preuves que le passage des États-Unis à cette méthode d'aggrégation élémentaire représente un pas dans la bonne direction.

- (2) Il ne semble exister aucune relation d'ordre claire entre les estimations britanniques au niveau de la *section* et les cibles correspondantes; par exemple, l'indice pour la section 1 est plus élevé que la cible L, tandis que l'indice pour la section 12 est plus faible que les indices superlatifs, etc. Toutefois, lorsque nous poursuivons l'aggrégation jusqu'aux niveaux du grand groupe et de l'ensemble des articles, les estimations commencent manifestement à s'approcher des indices superlatifs F ou T. (Dalen (1998) a noté un résultat similaire lors de l'aggrégation d'échantillons avec seuil d'inclusion.)

- (3) Si nous choisissons l'indice de Fisher comme cible, *même au niveau de la section*, la racine carrée de l'erreur quadratique moyenne de l'estimateur du Royaume-Uni est nettement plus faible que celle de l'estimateur des États-Unis. Étant donné la nature relativement restreinte du plan d'échantillonnage du Royaume-Uni, il n'est pas étonnant que l'estimateur de ce pays présente une variance plus faible, mais sa forme ne porterait pas à penser qu'il donne une approximation sans biais d'un indice de Fisher. Néanmoins, nos résultats laissent entendre que, du

moins pour une population d'achats tels que ceux utilisés dans l'étude, les méthodes par choix raisonné, à « force brute », du Royaume-Uni (et de nombreux autres pays) donnent de bons résultats.

Des résultats similaires ont été obtenus pour les paires successives d'années jusqu'à 1999–2000. La figure 6 donne la *moyenne géométrique* et l'indice de Fisher d'une année sur l'autre pour l'ensemble des articles pour cinq paires d'années, ainsi que les moyennes sur l'ensemble des échantillons des estimateurs américain et britannique correspondant. (Il convient de souligner la différence d'échelle entre la figure 6 et les figures 1 à 5.) Il est facile de voir que l'estimateur américain a tendance à suivre la *moyenne géométrique* de population. L'estimateur britannique, qui suit l'indice de Fisher, a tendance à surestimer les prix dans les années les plus récentes, bien qu'il s'approche nettement plus de l'indice de Fisher que de la *moyenne géométrique* de population. Il convient de souligner que nous avons utilisé des données sur les dépenses de plus en plus pétimées, à savoir les données pour 1995, pour l'échantillonnage et l'estimation. Il se peut que les données pétimées sur les dépenses aient une incidence plus grande sur les estimations britanniques que sur les estimations américaines, peut-être en nous menant à surestimer les articles représentatifs coûteux ou à nous concentrer sur certains groupes de magasins qui pratiquent des prix de plus en plus élevés.

Les résultats pour les catégories (« chaudes », etc.) étaient forts semblables pour les États-Unis relativement à la *moyenne géométrique* et ne sont pas présentés. La figure 7 donne la différence entre les estimations britanniques moyennes d'une année sur l'autre et l'indice de Fisher, pour chacune des quatre catégories. Son examen révèle que la tendance à la surestimation dans les années les plus récentes affecte les quatre catégories.

Dans l'ensemble, les estimateurs du Royaume-Uni fournissent de meilleures estimations de l'indice superlatif cible de Fisher que ceux des États-Unis. Le tableau 8 donne le ratio de la racine carrée de l'erreur quadratique moyenne du Royaume-Uni à celle des États-Unis, pour les cinq paires d'années, pour l'ensemble des articles, pour les groupes et pour les sections. Il contient quelques valeurs anormales, notamment dans les indices 1998–1999 où, pour la section 2 de « chaudes » et, par conséquent, pour la catégorie « chaudes » complète, les estimations britanniques sont appéciablement moins bonnes. Cependant, en général, les méthodes du Royaume-Uni produisent de nettement meilleures estimations. Cela est attribuable, en partie, à une structure d'échantillonnage plus stricte (principalement parce que l'échantillonnage par choix raisonné/avec seuil d'inclusion est sensiblement plus contraignant que la sélection aléatoire de l'ensemble d'articles qui peuvent entrer dans l'échantillon), qui produit, ce qui n'est pas

entendons l'exploitation d'une combinaison de deux facteurs qui jouent souvent un rôle dans l'établissement des prix et la construction des indices des prix. En premier lieu, les leaders du marché ont tendance à dicter le prix; par exemple, s'ils augmentent ou réduisent fortement les prix, leurs concurrents moins importants vendant des biens similaires peuvent penser qu'il est nécessaire ou justifié de suivre leur exemple. En deuxième lieu, même si la tendance des prix varie entre biens semblables, les principaux vendeurs domineront vraisemblablement l'indice des prix en raison du montant élevé des dépenses, autrement dit, à cause de leur pondération conséquemment élevée.

La méthode américaine est l'échantillonnage et l'estimation probabiliste, typiquement dpw , et celle de la méthode britannique est l'échantillonnage sélectif, en choisissant l'article ou la catégorie que l'on estime avoir le plus d'importance selon le montant des dépenses ou la quantité vendue. Les méthodes de formation des agrégats élémentaires diffèrent et les poids utilisés pour l'agrégation au Royaume-Uni sont estimés à un niveau un peu moins fin de détail aux étapes inférieures.

Le tableau 5 résume ce que l'on pourrait considérer comme étant les points forts et les points faibles des méthodes américaine et britannique. Par avantage de « force brute », que nous attribuons à l'approche britannique, nous

Tableau 4
Comparaison sommaire des méthodes appliquées aux États-Unis et au Royaume-Uni

É.-U.		R.-U.	
Enquête sur les dépenses	E_{95}^c	E_{95}^c, E_{95}^h	
des ménages			
Dépenses par point de			
vente/catégorie	Ménages (POPS) E_y^{jc}	Enquête auprès des magasins (ARL) E_y^{jh}	
Choix des catégories	2 ANE c/strate d'articles h/UPE l	2 articles représentatifs c/section h/région l <i>Les plus importants</i> (E_{95}^c / E_{95}^h)	
d'articles	$pplar$ (E_{95}^c / E_{95}^h)	8 points de vente j/UPE l	8 points de vente j'article représentatif <i>c</i> × région l – <i>cas</i> avec type de magasin k_i $E_{95}^h > 0$
Article dans un point de	1 article i/lfc ppi (E_y^{jc} / E_y^{jc})	1 variété i/lfc max [$q_{ji}^c q_{ji}^h + 1$]	
vente/une catégorie			
Indice élémentaire	$I_{y,y+1}^h = \prod_{i \in I_{hcs}^{(j)}} \left(\frac{p_{ijh}^{h+1}}{p_{ijh}^h} \right)$	$I_{hbc}^{jc} = \frac{1}{\sum_{j \in I_{hbc}} \frac{1}{p_{jhci}^{y+1}}} = \frac{1}{\sum_{i \in I_{jhc}} \frac{1}{p_{ijhc}^{y+1}}}$	$I_c = \sum_{j \in I_{hbc}} \sum_{i \in I_{jhc}} \frac{1}{p_{jhci}^{y+1}} \frac{1}{p_{ijhc}^{y+1}}$ $\hat{W}_{hbc} = f(\hat{E}_{jhc}, \hat{E}_{l,h})$
Niveau plus élevé	$\frac{\sum_{j \in I_{hcs}^{(j)}} \frac{1}{I_{y,y+1}^h} \sum_{i \in I_{hcs}^{(j)}} I_{ijh}^h}{\sum_{i \in I_{hcs}^{(j)}} I_{ijh}^h}$		
d'agrégation			

Tableau 5
Comparaison des approches américaine et britannique

Points forts		Points faibles	
É.-U.	Recueille plus d'information	Répétition éventuelle lors de la sélection	
	Plus grande utilisation de l'information	Ne tient pas compte de la stratification des magasins	
	Sait la théorie classique de l'échantillonnage	(c'est-à-dire, de la classification en chaînes)	
	Donne des estimateurs pondérés des estimations		
	régionales (UPE) au niveau le plus faible		
	Procédure opérationnelle plus normalisée		
R.-U.	S'appuie sur le principe de « force brute »	Ensemble disparate de coefficients de pondération	
	Stratification des points de vente	Inconstante dans le cas de l'agrégation des prix	
	L'enquête sur le terrain auprès des magasins	Estimateur non pondéré et en apparence arbitraire au	
	s'appuie sur diverses sources	niveau le plus faible.	

3. Nous échantillonons la variété i avec $\text{Max}\{q_i^*, q_i^{**}\}$.

Naturellement, ce processus requiert plus d'information que n'en posséderait un économiste de terrain à la première période (et, de nouveau, n'est pas utilisé dans la méthode d'échantillonnage américaine décrite plus haut), et peut être considéré comme offrant un substitut pour l'évaluation d'éré comme offrait un substitut pour l'évaluation relative des biens vendus.

Nota : Il est commode d'utiliser l'expression échantillonnage *maximizing* pour désigner la combinaison de la sélection d'un point de vente par *ease* comme en (b) et d'un article dans le magasin comme en (c).

3.4 Estimation au Royaume-Uni

Pour le Royaume-Uni, les agrégats élémentaires ont été calculés au moyen d'une formule de ratio des moyennes (RM) dans chaque cellule de classification croisée définie par la région, le type de magasin et l'article représentatif. Il s'agit fondamentalement d'une estimation non pondérée donnée, pour les magasins indépendants, par

$$\frac{1}{\sum_{i \in c, j \in k} p_{ijci}^{y+1}} = \frac{1}{\sum_{i \in c, j \in k} p_{ijhc}^y}$$

Dans le cas des magasins multiples, une version pondérée de la formule susmentionnée est utilisée avec les dépenses par type de magasin, estimées d'après l'ARL, qui fournissent les poids relatifs des magasins multiples à relevé des prix centralisé par opposition à non centralisé.

Un indice à l'échelle du pays pour les articles représentatifs c dans l'échantillon (agrégés sur les types de magasin k et les régions l) est alors calculé à l'aide d'un estimateur de type Laspeyres :

$$I_{y, y+1}^c = \frac{\sum_{i \in c} \sum_{j \in k} p_{ijhc}^y}{\sum_{i \in c} \sum_{j \in k} p_{ijci}^{y+1}}$$

où $c \in h$, et \tilde{w}_{lhc}^y est fondé sur E_{yh}^y provenant de l'ARI et E_{95}^{lh} provenant de la FES (l'utilisation de ces périodes fait en sorte que l'information utilisée est la même pour les États-Unis et le Royaume-Uni). Une agrégation supplémentaire (sur les articles représentatifs c) est faite en utilisant E_{yh}^{hc} , etc. provenant de la FES.

3.5 Comparaison

Le tableau 4 donne une comparaison sommaire des méthodes appliquées aux États-Unis et au Royaume-Uni que nous avons considérées. La caractéristique prédominante de

3 parmi les magasins indépendants (chaîne 8) prix centralisé (chaîne 4)

1 provenant d'un magasin multiple à relevé de

1. L'information sur le type de magasin est utilisée pour la stratification (et jouera un rôle dans l'estimation décrite plus loin). Cette information est disponible dans l'échantillon des États-Unis, mais est passée outre au profit de la méthode d'échantillonnage *dpi*.

2. Pour le Royaume-Uni, nous permettons de l'information sur la présence ou l'absence de l'article représentatif spécifique c (équivalent à l'ANE) dans la liste de magasins avant l'échantillonnage, tandis que pour les États-Unis, on ne connaît effectivement que l'existence d'un certain ANE dans la strate d'articles donnés. (Cela sous-entend des mises en correspondance multiples de catégories ANE à POPS, ce qui était typiquement le cas jusqu'à récemment aux États-Unis; la version courante des appariements de catégories ANE à TPOPS (téléphone point of purchase survey) est 1 à 1; autrement dit, une base de sondage de points de vente est construite pour chaque ANE.)

(c) Traditionnellement, pour chaque élément représentatif c , dans un magasin particulier, l'économiste de terrain sélectionne la variété i qu'il ou elle considère comme dominant les ventes, c'est-à-dire un échantillonnage au jugé de la variété achetée la plus systématiquement. Nous formalisons cela de la façon suivante :

1. Pour une paire magasin-article représentatif donnée (j, c), nous dressons la liste de toutes les variétés i .

2. Pour chaque variété, nous trouvons la quantité minimale $q_i^* = \text{Min}(q_i^y, q_i^{y+1})$ sur deux années.

3.3.2 Annual Retailing Inquiry (enquête-magasins)

L'objectif est d'obtenir des estimations des dépenses E_{lh} , par section et type de magasin. Cet objectif est considérablement plus général que l'obtention d'estimations point de vente (magasin) par ANE (article représentatif) visées par la POPs des États-Unis. Nous utilisons, pour construire les estimations ARI, pour chacun des 500 passages, les mêmes données que celles utilisées pour construire les estimations OOPS pour la simulation de l'IPC des États-Unis.

3.3.3 Échantillonnage des points de vente

La sélection des articles pour lesquels les prix doivent être relevés comprend les étapes suivantes :

- a) Un « échantillon au jugé » d'articles représentatifs c est sélectionné dans chaque section h . Dans la présente étude (uniquement pour permettre la simulation), dans chaque section, nous sélectionnons les deux articles représentatifs ayant les valeurs les plus grandes de E_{hc} . Il convient de souligner deux différences par rapport à l'étape (a) correspondant de la méthode américaine : i) la sélection est uniforme sur l'ensemble des régions ; ii) la sélection n'est pas aléatoire et, en particulier, ne permet pas la sélection répétée des articles représentatifs. (La sélection répétée peut avoir lieu dans la méthode américaine simulée, à cause de l'échantillonnage avec remise des ANE dans les strates d'articles.)
- b) Les économistes de terrain choisissent les magasins dans une localité particulière dans laquelle le prix d'un article représentatif doit être établi. Traditionnellement, cela se faisait par *easey*, après que l'économiste de terrain ait construit une base de sondage des magasins appropriés. Plus récemment, la sélection a été faite par échantillonnage *ppt*, où la mesure de taille est la superficie consacrée dans le magasin au type de biens que l'article représente-tatif représente. Les économistes de terrain ne tirent pas d'échantillons d'articles « dont le relevé des prix est centralisé » : dans le cas d'un très grand magasin multiple, le prix d'un article est relevé auprès du bureau central de ce magasin et est considéré comme représentatif du prix de l'article dans tous les magasins faisant partie du multiple. Dans la présente étude, nous avons procédé de la façon suivante : pour chaque région l et chaque article représentatif c , nous avons sélectionné huit magasins comme il suit :

- 4 parmi les magasins multiples à relevé de prix non centralisé (chaînes 1, 2, 3, 5, 6, 7)

$$W_{jhc}^{lhc} = \frac{E_{lc} E_{lh}}{E_{lh}^2} W_{jhc}^{lhc}$$

base; voir le *BLS Handbook of Methods* (2005).
Puis, les indices élémentaires sont agrégés en utilisant les dépenses estimées d'après la CEX conformément à la formule de Laspeyres, par exemple

$$I_{y,y+1}^h = \frac{\sum_l E_{lh}^{y+1}}{\sum_l E_{lh}^y}$$

pour obtenir l'indice pour une strate d'articles donnée h , sur l'ensemble des UPE.

3.3 Méthodes d'échantillonnage du Royaume-Uni

Au Royaume-Uni, comme aux États-Unis, la méthode d'estimation comporte la combinaison de trois composante : 1) une enquête-ménages, appelée Family Expenditure Survey (FES), pour estimer les montants consacrés à l'achat de divers groupes d'articles, 2) une enquête-magasins, appelée Annual Retailing Inquiry (ARI), pour obtenir des renseignements sur les dépenses par section et type de magasin et 3) une enquête auprès des points de vente des magasins, pour sélectionner des articles pour l'établissement des prix.

3.3.1 FES (enquête-ménages)

L'objectif est d'estimer les dépenses E_{lc} pour des articles représentatifs c , et les dépenses E_{lh} pour des machines par passage-machine entre les données pour la CEX des États-Unis et pour la FES du Royaume-Uni, de sorte que nous avons, de nouveau, 500 ensembles de données FES. Notons que le Royaume-Uni ne cherche pas à obtenir les estimations plus détaillées E_{lhc} que visent les États-Unis.

Les dépenses de plus haut niveau ont été estimées par simple sommation. Par exemple, le total, étendu à l'ensemble des UPE, dans un ANE donné c est estimé par $E_{95}^c = \sum_i E_{95}^{ic}$, etc. En tout, nous avons tiré 500 échantillons CEX, chacun produisant un ensemble correspondant d'estimations des dépenses.

3.1.2 POPS (enquête-ménages)

L'objectif de cette enquête est d'estimer la distribution des dépenses dans différents points de vente pour des catégories particulières de biens. Ces catégories pourraient être des ANE ou des groupes d'ANE; ici, nous supposons qu'il s'agit d'ANE. La TPOPS (Telephone Point of Purchase Survey) réelle des États-Unis est, comme son nom l'indique, réalisée par téléphone, selon un plan avec renouvellement de l'échantillon tous les quatre ans. Nous nous sommes efforcés, comme nous l'avons fait dans le cas de la CEX, de faire concorder les propriétés statistiques de notre procédure avec celles de la TPOPS réelle, mais il s'est avéré que faire correspondre les tailles d'échantillon dans notre fichier de 20 000 enregistrements nous aurait donné des fractions d'échantillonnage dans les UPE plus grandes qu'il n'était souhaitable. Par conséquent, nous avons réduit les tailles d'échantillon de moitié, de sorte que notre « imitation de POPS » devrait avoir une précision d'environ $1/\sqrt{2}$ celle de la TPOPS réelle. De nouveau, cette modification n'aura pas d'incidence sur les conclusions de l'étude, parce que nous avons utilisé des données identiques pour la construction de l'enquête britannique. Nous avons tiré des échantillons $s(p_i)$ de taille n_i par *casar* et procédé à l'estimation au moyen de l'estimateur à facteur

$$E_{ly}^c = \frac{N_y}{n_y} \sum_{i \in c(s(p_i))} E_{y^i}^c$$

Puisque les données de la POPS ont tendance à être plus à jour que celles de la CEX, nous choisissons y comme année de base de l'indice, 1995 pour 1995–1996, mais 1996 pour 1996–1997, etc. Nous avons exécuté 500 passages machines et obtenu 500 ensembles d'estimations, qui ont chacun été apparié à une réalisation de la CEX.

3.1.3 Échantillonnage des points de vente

Pour chaque année y , la sélection des articles pour lesquels les prix doivent être relevés comprend les étapes suivantes :

a) Pour chaque UPE i , et chacune des dix strates d'articles h , nous sélectionnons deux ANE c par échantillonnage avec probabilité proportionnelle à la taille avec remise ($pptar$), avec la mesure de la taille E_{95}^{ic} dérivée de la CEX.

b) Pour chaque ANE c sélectionné, nous tirons huit points de vente j par échantillonnage $pptar$, en utilisant comme mesure de taille les estimations des dépenses d'après la POPS E_y^{jc} . Donc, en tout, nous obtenons 160 paires ANE-point de vente par UPE, et un nombre total de 480, avec un certain degré de répétition éventuel.

c) Dans chaque groupe point de vente-ANE (j, c), nous « allons » (comme l'agent de terrain tirait littéralement) dans le point de vente et « dressons la liste » de tous les articles appartenant à l'ANE et des dépenses correspondantes à la première période E_y^{jic} et, au moyen de cette base de sondage dans le point de vente, nous sélectionnons un article par échantillonnage ppt .

Pour chaque article ainsi sélectionné, nous enregistrons les prix P_y^{jic} , $y = 1, 2$. Donc, nous notons que tous les aspects de l'échantillonnage des points de vente sont ppt avec remise, en fonction des estimations des dépenses provenant de l'une ou l'autre des deux enquêtes-ménages ou provenant directement du magasin sélectionné. De nouveau, nous avons exécuté 500 passages machines, chacun correspondant à une réalisation CEX/POPS unique.

3.2 Estimation aux États-Unis

Les « agrégats élémentaires » F_y^{y+1} , c'est-à-dire des estimations d'indice au niveau de l'UPE \times strate d'articles, sont les éléments à partir desquels est construit l'IPC. Dans la plupart des IPC partout dans le monde, les indices de niveau le plus faible sont des moyennes non pondérées d'une sorte ou l'autre, comme l'estimateur RM du Royaume-Uni décrit plus loin, et les données sur les dépenses ne sont utilisées que pour agréger ces indices à des niveaux plus élevés. Aux États-Unis, les indices élémentaires sont essentiellement des estimateurs d'Horvitz-Thomson fondés explicitement ou implicitement sur les estimations des dépenses provenant de la CEX ainsi que de la POPS. Ces États-Unis ont adopté la formule de la moyenne géométrique (voir l'annexe A), de sorte que les estimations à ce niveau prennent la forme

$$F_y^{y+1} = \prod_{i \in l, j \in h} \left(\frac{P_y^{jic}}{P_{y+1}^{jic}} \right)^{W_{jic}^{lh}}$$

où

$$s_{jic}^{lh} = \frac{\sum_{j \in l, c \in h, i \in c} W_{jic}^{lh}}{W_{jic}^{lh}}$$

dans la suite) pour simuler toutes les phases des opérations aux Etats-Unis et au Royaume-Uni.

3. Méthodes d'échantillonnage simulées

Les méthodes d'échantillonnage compliquées que nous avons utilisées pour simuler les approches adoptées aux

Etats-Unis et au Royaume-Uni sont modélisées d'après les pratiques respectives de ces deux pays. Ces pratiques évoluent au cours du temps et présentent même des variantes à un point particulier dans le temps. Notre objectif n'était pas de déterminer quel pays utilise la meilleure méthode, ni d'englober toutes les variantes. Nous cherchions plutôt à comparer deux modes distincts d'échantillonnage, en tenant compte de la gamme de complexités que postulent ces modes. Le lecteur que cela intéresse trouvera une description de la construction de l'IPC des Etats-Unis dans le *BLS Handbook of Methods* (2005), chapitre 17. Pour l'indice des prix de détail du Royaume-Uni, nous sommes fondés sur le document intitulé *The Retail Prices Index Technical Manual* (1998). Une description des pratiques plus récentes adoptées au Royaume-Uni peut être consultée dans le *Consumer Price Indexes Technical Manual* (2005).

3.1 Méthodes d'échantillonnage aux Etats-Unis

Nous commencerons par décrire les méthodes d'échantillonnage appliquées aux Etats-Unis, qui nécessitent trois enquêtes avec échantillonnage probabiliste, à savoir 1) une enquête-ménages, appelée *Consumer Expenditure Survey* (CEX), pour estimer la répartition des dépenses des ménages entre diverses catégories de biens, 2) une deuxième enquête-ménages, appelée *Point of Purchase Survey* (POPS) pour estimer, dans chaque groupe d'articles, les montants relatifs des dépenses dans divers points de vente, et 3) une enquête auprès des points de vente, grâce à laquelle sont sélectionnés des articles individuels dont le prix est relevé. Dans les trois cas, l'échantillonnage pour la simulation est aléatoire avec remise (quoique l'échantillonnage employé en pratique soit nettement plus compliqué). Les deux premières enquêtes sont fondées sur des échantillons aléatoires simples et la

3.1.1 CEX (enquête-ménages)

ANE.

L'objectif est d'estimer E_{lc} , c'est-à-dire les dépenses brutes des ménages au titre de l'ANE c dans l'UPE l . Nous avons procédé à un échantillonnage aléatoire simple avec remise (*easy*) à partir du fichier décrit plus haut, dans les UPE, de manière à obtenir des estimations à facteur d'extension sans biais

$$E_{lc}^{95} = \frac{N_{95}^l}{N_{95}^{lc}} \sum_{j \in c(s(xl))} n_{xl}^{j \in c(s(xl))} E_{jlc}^{95}$$

où $E_{jlc}^{95} = q_{jlc}^{95} P_{jlc}^{95} N_{95}^{lc}$ est la taille de population (nombre d'enregistrements pour l'UPE l dans 1995–1996) et $n_{xl}^{j \in c(s(xl))}$ est la taille de l'échantillon CEX $s(xl)$ dans l'UPE l , choisies de façon à concorder avec les tailles réelles d'échantillon de la CEX américaine et d'obtenir des coefficients de variation des estimations s'approchant de ceux obtenus dans le cas de la CEX américaine réelle; le x dans $s(xl)$ et dans $n_{xl}^{j \in c(s(xl))}$ sert simplement à faire la distinction entre l'enquête CEX et l'enquête POPS (dont la notation correspondante est « p »; voir plus loin) ou l'enquête sur les prix. Cette « imitation de la CEX » est une version simplifiée de l'enquête réelle. Notre méthodologie reposait sur l'hypothèse tacite que tous les clients d'un point de vente donné achètent les articles dans les mêmes proportions; elle ne tenait pas compte de l'erreur de mesure inévitable dans toute enquête réelle sur les dépenses, et (pour 1995–1996) elle était trop récente : plusieurs années aux enquêtes auprès des points de vente pour lesquelles elles sont utilisées. Toutefois, puisque les « données des ménages » recueillies ont aussi été utilisées dans les méthodes correspondantes du Royaume-Uni (voir plus loin), la version simplifiée suffit pour la comparaison souhaitée des méthodologies.

Tableau 3 Structure de population de l'« univers des céréales » : points de vente

R.-U.	É.-U.	Symbolique
Région	Unité	primaire 3
Type de magasin :	Indépendant	k
Multiple : $\left\{ \begin{array}{l} \text{Relevé central} \\ \text{Non central} \end{array} \right.$	Chaîne 8	
	Chaîne 4	
	Chânes 1 à 3 ; 5 à 7	
	Point de vente	~300
Magasin		j

En ce qui concerne les mérites relatifs des méthodes de base utilisées dans les deux pays.

Tableau 2
Structure de population de l'« univers des céréales » : articles

Nombre de groupes		Symbole		E.-U.	
Grand groupe		Catégorie de dépenses		Srate d'articles	
Section		Article de niveau		Article	
Variété		d'entrée (ANE)		326	
				29	
				10	
				4	

Les quatre groupes de population « naturels » décrits plus haut, qui sont appelés « grands groupes » au Royaume-Uni et « catégories de dépenses » aux États-Unis, ont été divisés en sous-groupes moins agréés. En pratique, les sous-groupes seraient définis en fonction de types de produits. L'une des raisons, outre tout intérêt intrinsèque que l'on pourrait avoir pour ces produits, est que les sous-groupes ainsi formés ont tendance à être homogènes en ce qui a trait aux tendances des prix. Aux fins de la présente étude par simulation, nous définissons par conséquent les sous-groupes de la façon suivante :

- 1) Nous avons calculé la variation de prix de long terme pour chacun des 326 articles compris dans les données de base, en utilisant les indices de valeur unitaire pour les articles (sur l'ensemble des points de vente) pour 2000 comparativement à 1995.
- 2) Nous avons ajouté un bruit à ces indices, tiré les articles dans chaque grand groupe en fonction de leur valeur de l'indice perturbé, et regroupé les articles adjacents. Le regroupement des articles dont les indices de long terme étaient proches avait pour but de rendre les sous-groupes homogènes, et j'ajout d'un bruit à été fait de sorte que l'homogénéité soit raisonnablement imparfaite.

Le tableau 2 donne la structure des articles de la population qui a été constituée, y compris la nomenclature utilisée dans les deux pays, le nombre de groupes à chaque niveau d'agrégation et le symbole correspondant à chaque niveau de classification utilisé dans le présent article. L'« article représentatif » est le niveau d'agrégation le plus faible auquel un indice est produit aux Royaume-Uni. Il correspond à l'article d'entrée au ANE (en anglais *Entry Level Item* ou ELI) des États-Unis, qui est en fait un ensemble d'articles similaires ou connexes. Aux États-Unis, les indices sont produits pour les catégories obtenues au niveau directement supérieur d'agrégation, c'est-à-dire le niveau de la « strate d'articles », mais ces catégories sont encore subdivisées en fonction des régions géographiques dans lesquelles les articles sont vendus. Notons qu'il existe 2 ou 3 strates d'articles/sections h dans une catégorie/un grand groupe C , 3 ANE/articles représentatifs c par strate d'articles/section h (sauf dans un cas où il y en a 2), et 10 d'articles/variétés i dans chaque ANE/article représentatif c . (Nota : une catégorie réelle du Royaume-Uni pourrait être plus grande ou plus petite que la catégorie correspondante des États-Unis; par exemple, en règle générale, l'ANE comprend probablement plus de sortes d'articles spécifiques que l'article représentatif. Nous avons donc dû forcer l'équivalence pour assurer que la même quantité d'information soit utilisée dans chaque approche. Cet ajustement n'aura pas d'incidence sur nos conclusions.)

En plus de la structure fondée sur les articles, chaque population de transactions possède une structure « spatiale » caractéristique de l'endroit où a été vendu un article. Cette structure est résumée au tableau 3. Les points de vente qui recoupent les trois unités primaires d'échantillonnage géographique des États-Unis à partir desquelles les données sur les céréales ont été recueillies. (Dans la terminologie du Royaume-Uni, les chaînes sont appelées « *magasins multiples* ».) Les points de vente appartenant à une chaîne donnée ont un propriétaire commun, à l'exception de la « chaîne 8 », qui est un groupe « fourre-tout » composé des points de vente n'appartenant pas à une grande chaîne (il pourrait y avoir certaines « mini-chaînes »). Lors de l'appartenance de cette « structure basée sur les chaînes » à la classification des points de vente utilisée pour l'échantillon-mage au Royaume-Uni, la chaîne 8 a été considérée comme un ensemble de « magasins indépendants » (le terme utilisé pour désigner les magasins appartenant à un propriétaire indépendant au Royaume-Uni). La chaîne 4, qui semble présenter la plus grande homogénéité des prix sur l'ensemble des points de vente, a été considérée comme un « magasin multiple pour lequel le relevé des prix est centralisé » (centra lly collected multiple), expression utilisée au Royaume-Uni pour les groupes de points de vente dont les prix sont contrôlés centralement. Chaque chaîne restante a été considérée comme une chaîne dont le relevé des prix n'est pas centralisé. Les méthodes de collecte et d'estimation pour ces trois types de chaînes sont données plus loin dans la description des méthodes appliquées au Royaume-Uni.

Donc, la population est constituée de $N \approx 20\,000$ enregistrements pour les indices de 1995–1996, chaque enregistrement représentant l'achat d'un article i dans un point de vente j . Sont reliés à chaque article/point de vente son UPE/région l , sa chaîne/son type de magasin k , l'ANE/point de vente/magasin j , l'article/la variété i , l'ANE/article représentatif c , la strate d'articles/section h , la catégorie de dépenses/le grand groupe C , et p_y , q_y , p_{y+1} , et q_{y+1} , les prix et quantités (en onces) des articles vendus (en février des) deux années en question. Nous avons utilisé ce fichier de population (appelé simplement « le fichier »

Tableau 1
Indices directs et enchaînés pour 1995 à 2000

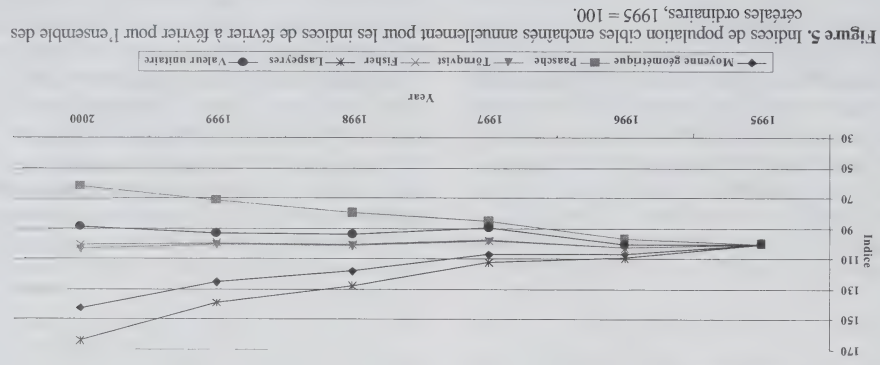
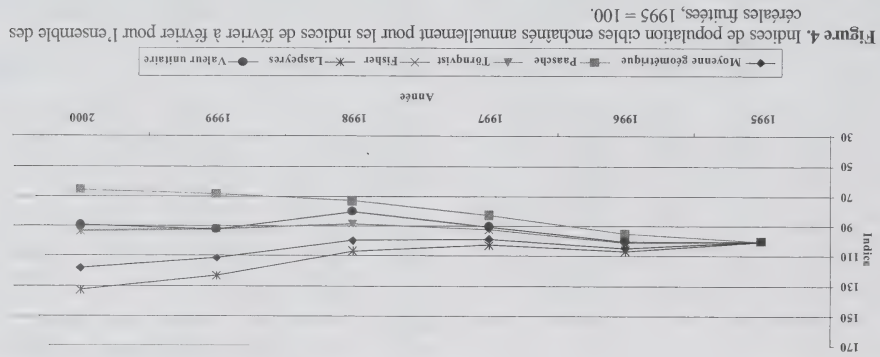
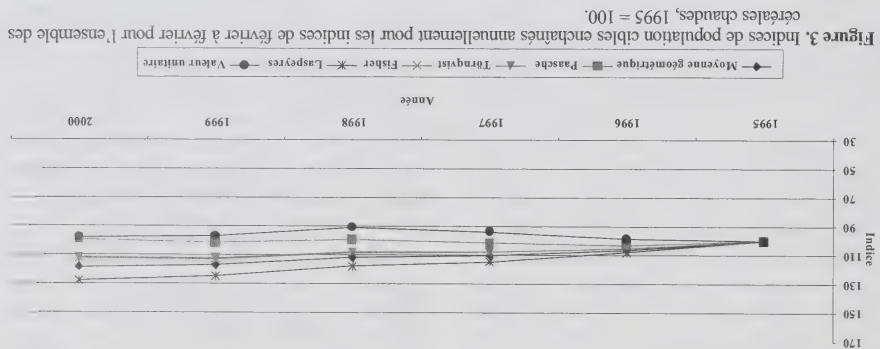
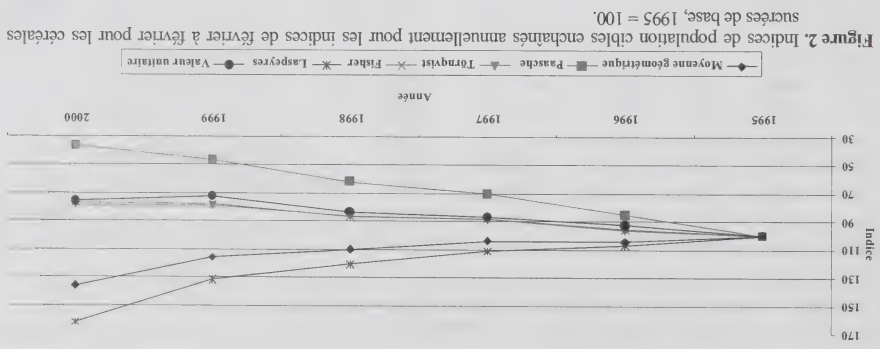
Chaudes	Direct	Moyenne géométrique						
		Paasche	Törnqvist	Fisher	Laspeyres	Unitaire	RM	Moyenne géométrique +
Sucrées	Direct	1,1176	1,0253	1,0847	1,1569	0,9576	1,1192	1,0949
	Enchaîné, ensemble des articles	1,1801	0,9874	1,1159	1,1216	1,2742	0,9453	1,1128
	Enchaîné, articles de base	1,1804	0,9865	1,1160	1,1221	1,2763	0,9759	1,1151
Fruitées	Direct	0,8855	0,6739	0,7913	0,7898	0,9257	0,7417	0,8702
	Enchaîné, ensemble des articles	1,3341	0,3825	0,7925	0,7771	1,5786	0,7506	0,9010
	Enchaîné, articles de base	1,3591	0,3661	0,7849	0,7704	1,6212	0,7585	0,8894
Fruitées	Direct	0,9716	0,8676	0,9319	0,9296	0,9960	0,8932	0,9726
	Enchaîné, ensemble des articles	1,2202	0,6849	0,9661	0,9696	1,3728	0,9308	1,0165
	Enchaîné, articles de base	1,1808	0,6557	0,9320	0,9328	1,3269	0,8950	0,9820
Ordinaires	Direct	1,0811	0,8641	1,0045	0,9816	1,1150	0,8554	1,0511
	Enchaîné, ensemble des articles	1,3969	0,6330	1,0333	1,0053	1,5965	0,8935	1,0572
	Enchaîné, articles de base	1,4234	0,6175	1,0353	1,0054	1,6370	0,8879	1,0571

Tableau 1
Indices directs et enchaînés pour 1995 à 2000

Chaudes	Direct	Moyenne géométrique*						
		Paasche	Törnqvist	Fisher	Laspeyres	Unitaire	RM	Moyenne géométrique +
Enchaîné, ensemble des articles	1,1176	1,0253	1,0847	1,0891	1,1569	0,9576	1,1192	1,0949
Enchaîné, ensemble des articles	1,1801	0,9874	1,1159	1,1216	1,2742	0,9453	1,1395	1,1128
	1,1804	0,9865	1,1160	1,1221	1,2763	0,9759	1,1374	1,1151
Direct	0,8855	0,6739	0,7913	0,7898	0,9257	0,7417	0,8817	0,8702
	1,3341	0,3825	0,7925	0,7771	1,5786	0,7506	0,9124	0,9010
	1,3591	0,3661	0,7849	0,7704	1,6212	0,7585	0,8984	0,8894
Fruitées	Direct	0,9716	0,8676	0,9319	0,9296	0,9960	0,8932	0,9726
	1,2202	0,6849	0,9661	0,9696	1,3728	0,9308	1,0263	1,0165
	1,1808	0,6557	0,9320	0,9328	1,3269	0,8950	0,9935	0,9820
Ordinaires	Direct	1,0811	0,8641	1,0045	0,9816	1,1150	0,8554	1,0511
	1,3969	0,6330	1,0333	1,0053	1,5965	0,8935	1,0642	1,0572
	1,4234	0,6175	1,0353	1,0054	1,6370	0,8879	1,0653	1,0571

* Pondérée par les dépenses à la période de base.
+ Non pondérée.

D'après les examens préliminaires, et pour simplifier, nous limitons nos investigations plus approfondies aux données sur les articles de base. Afin d'étudier l'exactitude relative des échantillonnages probabiliste et par choix raisonné, tels qu'ils sont appliqués en pratique pour établir les IPC, nous nous sommes efforcés d'approximer les plans d'échantillonnage utilisés aux États-Unis et au Royaume-Uni, qui représentent l'échantillonnage probabiliste et l'échantillonnage par choix raisonné, respectivement. Dans les deux cas, nous avons eu la chance de disposer d'information détaillée sur les processus d'enquête complexes grâce à des manuels et à des contacts avec les organismes respectifs. L'idée fondamentale consistait à procéder à l'échantillonnage répété d'une population donnée, par exemple, les transactions de base effectuées en 1995 et en 1996. Chaque « passage-machine » était une combinaison d'activités d'échantillonnage et d'estimation exécutées conformément aux méthodes d'un pays ou de l'autre. Il ne faut pas oublier que nous voulions comparer les mérites des *méthodologies* et non évaluer le succès avec lequel les États-Unis et le Royaume-Uni estiment les paramètres de leur population cible.



articles (de base et autres). Au cours de la période de 1995 à 2000, il y avait 326 articles de base et, en tout, 848 articles distincts.

Les valeurs des indices annuels de population sont représentées aux figures 1 à 5. La figure 1 donne les valeurs de l'indice $I_{y,y+1}$ pour les céréales sucrées basées sur l'ensemble

des articles vendus dans les magasins durant les années y et $y+1$, pour (février de) $y = 1995, \dots, 1999$ (l'indice « d'ensemble »). Les valeurs sont présentées pour cinq indices, y compris l'indice de Paasche P et, en raison de son intérêt théorique, un indice de valeur unitaire U , le ratio des

prix moyens pondérés par les quantités, la moyenne étant calculée sur tous les types d'articles et tous les points de vente. La figure 2 donne les résultats des mêmes calculs, mais en se limitant aux articles « de base ». Les figures 1 et 2 sont presque identiques et la ressemblance entre les indices calculés en utilisant tous les articles (indice d'ensemble) et ceux obtenus en utilisant uniquement les articles de base est vérifiée pour les autres catégories de céréales également. Les figures 3 à 5 donnent les résultats relatifs aux indices calculés pour les articles de base dans le cas des céréales chaudes, pour les articles et ordinaires. Pour tout indice, les figures révèlent d'importants écarts entre les catégories de céréales. Les tendances des prix des quatre grands groupes sont assez différentes : celle de C est à la hausse, celle de S est fortement à la baisse, celle de F est moyennement à la baisse et celle de O est moyennement à la hausse.

Le tableau 1 donne les indices directs de long terme articles « et pour les « articles de base ». (« L'ensemble des données comprend les articles/points de vente pour lesquels les quantités vendues sont positives les deux années.) De nouveau, l'écart est très faible entre les valeurs obtenues pour les « articles de base » et « l'ensemble des articles », mais prononcé entre les catégories de céréales. Les résultats enchaînés et directs sont proches des indices superlatifs, mais ont tendance à diverger dans le cas de la *moyenne géométrique*, et des indices de Laspeyres et de Paasche. Les

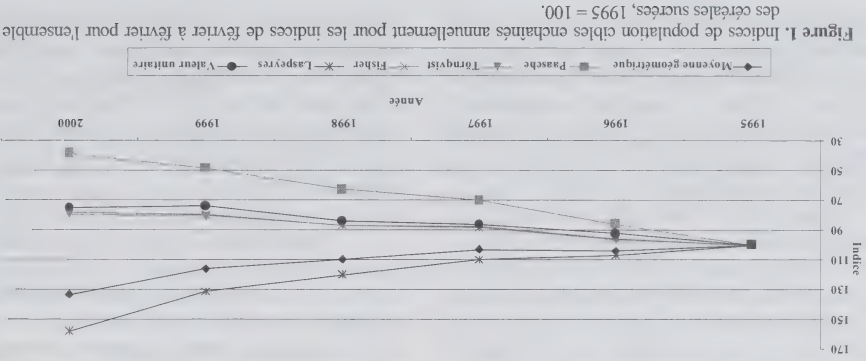


Figure 1. Indices de population cibles enchaînés annuellement pour les indices de février à février pour l'ensemble des céréales sucrées, 1995 = 100.

À part une certaine oscillation de la position de l'indice de valeur unitaire, nous observons un classement manifeste des indices en fonction de la formule, le même pour toutes les catégories de céréales, qui peut se résumer comme suit : (1) Les indices superlatifs diffèrent assez peu les uns des autres, résultat qui mérite d'être souligné étant donné le degré de variabilité due aux « ventes » des prix relatifs et des quantités liés aux combinaisons article-point de vente. (2) Les indices non superlatifs classiques diffèrent fortement les uns des autres et des indices superlatifs, la *moyenne géométrique*, pondérée par les dépenses à la première période, étant beaucoup plus élevée que les indices superlatifs, l'indice de Laspeyres étant encore plus élevé et l'indice de Paasche étant (sans surprise) nettement plus faible. Ces résultats donnent à penser que, dans l'univers des céréales, non seulement la quantité, mais aussi la part des dépenses, a tendance à diminuer à la période 2 pour un article dont le prix augmente fortement durant cette période. (3) L'indice de valeur unitaire est faible également, mais, sauf dans le cas des céréales chaudes, s'approche davantage des indices superlatifs que les indices non superlatifs classiques. (4) À la lumière de faits présentés plus loin dans l'article, et à la recommandation fondée sur les quantités dans ce tableau (mais pas dans les figures) : l'indice de *Duroi*, qui est un simple ratio des prix moyens (RM) – voir l'annexe A, et une *moyenne géométrique non pondérée* (c'est-à-dire pondérée par une consommation) au niveau élémentaire. Les résultats sont surprenants : en ce qui a trait à l'approximation des indices superlatifs, ils donnent d'aussi bons, voire de meilleurs, résultats que les indices non superlatifs basés sur les quantités traditionnelles, à peu près à égalité avec l'indice de valeur unitaire.

indices de prix globaux couvrant une vaste gamme hétérogène de produits, nous arriverons peut-être à dégager d'importants éclaircissements quant aux effets de diverses méthodes d'échantillonnage et au comportement d'estimateurs particuliers.

L'ensemble de données portant sur une période de six années nous a permis d'établir des tendances des prix d'assez long terme. Afin que le volume de données demeure raisonnable et pour éviter les complications de la saisonnalité, nous nous sommes limités aux données de février. Pour février de l'année y , pour chaque article (c) est-à-dire chaque combinaison particulière k de marque, type, format) dans un point de vente particulier, nous avons combiné les données sur les prix et les quantités pour une période de quatre semaines t en un prix et une quantité uniques pour un mois, en utilisant la somme des quantités vendues durant le mois comme valeur

$$q_{t \in Feb, y}^k = \sum_{t \in Feb, y} q_t^k \quad \text{et la quantité et la valeur unitaire } p_{t \in Feb, y}^k = \sum_{t \in Feb, y} p_t^k / \sum_{t \in Feb, y} q_t^k \text{ comme prix. Les valeurs unitaires calculées sur de courtes périodes (par exemple, un mois) correspondant à un article particulier. L'utilisation des valeurs unitaires lisse les données et les réduit à un volume plus raisonnable; pour une discussion des circonstances sous lesquelles l'utilisation des valeurs unitaires est ou non appropriée, voir Balk (1999).$$

Pour la présente étude, la population de céréales pour petit déjeuner a été subdivisée en quatre groupes :

1. céréales chaudes (C)
2. céréales « sucrées » (S)
3. céréales « fruitées » (F)
4. céréales « ordinaires » (O), c'est-à-dire les céréales froides ne rentrant pas dans les catégories (2) et (3).

Pour chaque groupe, pour chaque paire d'années successives, nous avons calculé les indices superlatifs et non superlatifs en utilisant les combinaisons article-point de vente disponibles les deux années. En pratique, il est généralement difficile d'obtenir des appartements parfaits de période en période, et il est important de trouver des moyens de faire face à ce problème en découvrant des substituts pour les produits originaux ou d'autres façons; la présente étude ne tient pas compte de ce problème particulier.

Nous avons calculé les indices de long terme (1995 à 2000) directement, ainsi que par enchaînement des indices annuels. En outre, nous avons calculé les indices fondés sur les articles « de base », c'est-à-dire ceux disponibles chaque année des six années. D'une année sur l'autre, les articles de base représentaient entre 53 % et 61 % des articles disponibles durant une année particulière pour la comparaison d'une année à l'autre; les dépenses de base représentaient entre 83 % et 91 % des dépenses totales au titre de tous les

Fisher ou de Törnqvist), même si la formule de l'estimateur ne ressemble pas à l'une des formules d'indice superlatif de population. Compte tenu de la forme des agrégats élémentaires utilisés aux États-Unis – moyenne géométrique – et du fait que des travaux de recherche antérieurs (Dorman, Leaver et Lent 1999) ont indiqué que le niveau le plus faible d'estimation peut avoir un effet important, la moyenne géométrique pondérée sera incluse parmi les indices cibles possibles. Les cibles pour un domaine particulier sont calculées en se basant sur les prix et sur les quantités de tous les articles compris dans le domaine, conformément aux formules qui figurent à l'annexe A (aggrégation à un degré des prix et des quantités).

Nota : Ces formules donnent l'illusion d'être simples, mais nécessitent la notation de la section 3 pour leur développement complet. Donc, dans une formule telle que celle de l'indice de Fisher F (que nous choisissons comme cible dans le corps de l'étude aux sections 2 à 4), « i » représente un article i appartenant à une petite catégorie c (un « ANB » [article de niveau d'entrée] ou « article représentatif » – voir la section 3), où c est elle-même un sous-ensemble d'une catégorie plus grande. En outre, l'article i est vendu dans un point de vente particulier j , classifié dans une région géographique échantillonnée particulière, dans une région géographique échantillonnée particulière, l'unité primaire d'échantillonnage (*upe*) l . Donc, dans le cas de l'indice de population global, l'expression pour une somme $\sum_{j \in (k, l)} \sum_{c \in h} \sum_{i \in (j, c)} \sum_{k=1}^4$ est en fait une abréviation de $\sum_{k=1}^4 \sum_{j \in (k, l)} \sum_{c \in h} \sum_{i \in (j, c)} \sum_{k=1}^4$. Brevement, il existe des sommes et des produits sur la totalité des articles dans la population. Les indices de population pour les diverses catégories C , etc.

2. La population pour l'étude principale

La source des données sur lesquelles porte l'étude est un ensemble de données scannées sur les céréales pour petit déjeuner couvrant la période de 1995 à 2000 dans trois zones séparées, mais contiguës, d'une grande région métropolitaine. Le U.S. Bureau of Labor Statistics a acheté l'ensemble de données à la société A.C. Nielsen en vue de déterminer s'il est possible d'intégrer des données scannées dans l'IPC des États-Unis; voir Richardson (2000).

Des « populations » artificielles ont été tirées à partir de ces données par la méthode décrite plus loin. Donc, l'étude a pour champ d'observation un univers en apparence limité, celui des céréales, dans un domaine géographique relative-ment restreint. Toutefois, même cet univers limité permet d'observer des tendances des prix assez divergentes au cours de la période de six ans. Donc, bien que nous ne puissions pas généraliser nos résultats, de façon simple, aux

d'une mesure préalable et équitable, d'évaluer la justesse de chaque estimation par rapport à un indice cible connu. Le présent article porte sur deux études, une grande étude principale et une étude de suivi secondaire plus petite. L'étude principale est décrite aux sections 2 à 4. La section 2 donne la construction de la population cible. La section 3 expose les méthodologies des « États-Unis » et du « Royaume-Uni », et fournit les renseignements sur les simulations. Aucun effort n'est fait pour évaluer les coûts relatifs (ce qui nous écarte de la situation idéale), mais les approches concurrentes sont rendues aussi égales que possible en ce qui a trait à l'information utilisée. Les résultats, qui donnent l'avantage à l'approche britannique, sont présentés à la section 4.

L'étude de suivi, décrite à la section 5, a pour but d'essayer de dégager les effets des diverses composantes des deux approches, en particulier la méthode d'échantillonnage et la formule de l'indice élémentaire. La section 6 comprend un résumé final et une discussion.

Note sur les indices cibles. La littérature sur les indices de prix contient des myriades de formules pour calculer la variation des prix d'une période à une autre. Divers indices sont comparables avec différentes hypothèses quant au comportement d'achat du consommateur « moyen » en réponse à la variation des prix. Les indices à « panier de consommation fixe », les formules fréquemment employées de Laspeyres et celles, moins fréquemment utilisées, de Paasche, sont comparables avec l'hypothèse selon laquelle les consommateurs continuent d'acheter les mêmes articles en même quantité quelle que soit la variation des prix relatifs. L'indice de Laspeyres projette les quantités de la période 1 (« période de base ») sur la période 2 (« période de base ») tandis que celui de Paasche applique les quantités de la période 2 à la période 1. L'indice géométrique (ou « de Jevons » ou « moyenne géométrique »), habituellement pondéré par les parts des dépenses à la période de base, suppose que le consommateur ajuste les quantités qu'il achète de telle façon que la part des dépenses pour chaque article demeure constante au cours du temps. Les formules des indices « superlatifs » de Fisher, de Tornqvist et de Walsh, qui reposent sur des données sur les quantités (ou parts des dépenses) pour les deux périodes, ne nécessitent pas ces hypothèses. Les formules de base, avec les exposants γ représentant la période de base, $\gamma + 1$, la période courante et 1 , l'article acheté, sont

Le débat concernant l'indice d'ensemble cible se résume habituellement à choisir entre l'indice de Laspeyres et l'un des indices superlatifs. La plupart des pays optent pour un indice cible de Laspeyres, mais de solides arguments peuvent être présentés (Diewert 1997) en faveur du choix d'un indice superlatif comme cible (habituellement celui de

justesse relative de diverses estimations basées sur un échantillon par rapport à la « valeur réelle »? De surcroît, dans la plupart des cas, on ne dispose même pas d'information échantillonnale pour l'un des éléments clés de l'indice de population, à savoir les *quantités* d'articles vendus, de sorte que même la construction d'une population artificielle à partir de données d'échantillon en vue d'une évaluation n'a pas été possible.

La disponibilité relativement récente de données *scanned*, aux États-Unis et ailleurs, offre une occasion sans précédent d'évaluer les approches d'échantillonnage et les estimateurs. Ces données comprennent les prix *et* les quantités, habituellement sur une base hebdomadaire, de *tous* les articles vendus dans une catégorie donnée dans un grand échantillon de points de vente dotés de scanners. Ce genre de données peuvent être utilisées pour construire des populations réalistes de transactions pour lesquelles l'indice des prix réel est connu. Nous pouvons alors utiliser diverses méthodes pour échantillonner cette population, construire différentes estimations d'indice d'intérêt et comparer les résultats aux paramètres de population connus. Une étude de ce type, décrite par de Haan, Opperdoes et Schun (1999), semble indiquer que l'« échantillonage avec seuil d'inclusion » (*coefficient sampling*) c'est-à-dire l'échantillonnage des quelques articles les plus importants (en ce qui a trait aux recettes générées) dans la population, donne de meilleurs résultats que deux grandes approches probabilistes, à savoir l'échantillonnage aléatoire simple (*eas*) et l'échantillonnage avec probabilité proportionnelle à la taille (*pda*) (où la mesure de taille est, de nouveau, les recettes).

L'une des difficultés que pose toute étude faisant ce genre de comparaison est la nécessité de maintenir des « règles du jeu équitables ». Si l'une des méthodes d'échantillonnage s'appuie, par exemple, sur des données (de population) susceptibles, en réalité, ne pas être disponibles en pratique, tandis qu'une autre ne le fait pas, la comparaison des méthodes est sérieusement minée. De même, si une méthode ne fournit qu'un seul échantillon ou quelques échantillons, et qu'une autre en fournit des milliers, des précautions particulières doivent être prises pour les comparer; en effet, ce genre de comparaison pourrait nécessiter d'importantes restrictions. Étant donné la complexité des méthodes d'échantillonnage et d'estimation utilisées pour le calcul des indices de prix, il n'est pas étonnant que ces difficultés et de nombreuses autres compliquent les expériences conçues en vue de comparer diverses méthodes. Idéalement, pour comparer les approches, disons, de deux pays, nous imiterions entièrement le processus complexe d'échantillonnage et d'estimation de chacun et nous évaluerions les coûts. Le même budget serait affecté aux deux processus et nous serions capables, au moyen

Plans de sondage pour les indices de la consommation

Alan H. Dorfman, Janice Lent, Sylvia G. Leaver et Edward Wegman¹

Résumé

L'échantillonnage en vue d'estimer un indice des prix à la consommation (IPC) est assez compliqué et requiert généralement la combinaison de données provenant d'au moins deux enquêtes, l'une dominant les prix et l'autre, la pondération par les dépenses. Deux approches fondamentalement différentes du processus d'échantillonnage – l'échantillonnage probabiliste et l'échantillonnage par choix raisonné – ont été vivement recommandées et sont utilisées par divers pays en vue de recueillir les données sur les prix. En construisant un petit « univers » d'achats et de prix à partir de données scannées sur les cartes à puce, puis en simulant diverses méthodes d'échantillonnage et d'estimation, nous comparons les résultats de deux approches du plan de sondage et de l'estimation, à savoir l'approche probabiliste adoptée aux États-Unis et l'approche par choix raisonné adoptée au Royaume-Uni. Pour la même quantité d'information recueillie, mais avec l'utilisation d'estimateurs différents, les méthodes du Royaume-Uni semblent offrir une meilleure exactitude globale du ciblage d'un indice superlatif des prix à la consommation basé sur la population.

Mots clés : Indice élémentaire; échantillonnage avec probabilité proportionnelle à la taille; échantillonnage par choix raisonné; données scannées; indice superlatif.

1. Contexte

Du début à la fin, l'échantillonnage en vue d'établir un indice des prix à la consommation (IPC) représente l'une des entreprises d'échantillonnage les plus compliquées. La population cible est difficile à définir, le domaine approprié des articles est débattu, les définitions des ingrédients bruts, c'est-à-dire les prix, les quantités et les articles, sont ambiguës et mises en doute. L'estimateur ultime – l'estimateur d'ensemble – repose sur des données provenant d'au moins deux enquêtes, l'une dominant les prix et l'autre, les « pondérations ». Sous le niveau des articles comme des « strates d'articles », c'est-à-dire des groupes d'articles dont il est supposé que le mouvement des prix est homogène, il n'existe habituellement pas de moyen rentable de tenir à jour les poids d'échantillonnage. Le débat se poursuit donc quant au choix approprié parmi divers estimateurs simples de la variation du prix pour les catégories d'articles, c'est-à-dire les « indices élémentaires ». La méthode appropriée d'agrégation de ces variations des prix, au moyen des pondérations, fait également l'objet de débats.

On distingue deux grandes approches de l'échantillonnage en vue de relever les prix : l'échantillonnage probabiliste et l'échantillonnage par choix raisonné, ou échantillonnage au jugé. L'approche d'échantillonnage la plus généralement reconnue consiste habituellement à injecter un élément aléatoire dans le processus de sondage et à se fonder sur cet élément aléatoire pour faire des inférences au sujet des caractéristiques de population que l'on veut étudier, c'est-à-dire un échantillonnage probabiliste ou « basé sur un plan de sondage »; voir, par exemple, Särndal, Swensson et

L'aspect intéressant est que fort peu d'évaluations de l'exactitude relative des diverses approches d'échantillonnage ont été faites. En réalité, il n'est pas certain que ce genre d'évaluation soit réalisable. Même pour les pays les plus petits, l'indice des prix sous-jacent calculé en se basant sur la population totale d'articles, ou indice des prix de population, englobe un si grand nombre de transactions portant sur un si grand nombre d'articles dans un si grand nombre d'emplacements qu'il est inaccessible. De surcroît, la population d'articles sur le marché varie constamment, ce qui complique l'application des formules d'indice de population classiques. Alors, comment peut-on juger de la

- Félix-Medina et Monjardin : Combinaison de l'échantillonnage par dépistage de liens et en grappes
- Heckathorn, D.D. (1994). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.
- Seber, G.A.F. (1982). *The Estimation of Animal Abundance*. 2^{ème} édition. London : Griffin.
- Snijders, T.A.B. (1992). Estimation on the basis of snowball samples: How to weight? *Bulletin de Méthodologie Sociologique*, 36, 59-70.
- Snijders, T.A.B. (1992). Rare populations, hidden populations, and link-tracing designs: What and why? *Bulletin de Méthodologie Sociologique*, 36, 34-58.
- Thompson, S.K., et Frank, O. (2000). Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépistage de liens. *Techniques d'enquête*, 26, 99-112.
- Evans, M.A., Kim, H.-M., et O'Brien, T.E. (1996). An application of profile-likelihood based confidence interval to capture-recapture estimators. *Journal of Agricultural, Biological, and Environmental Statistics*, 1, 131-140.
- Félix-Medina, M.H., et Thompson, S.K. (2004). Combining cluster sampling and link-tracing sampling to estimate the size of hidden populations. *Journal of Official Statistics*, 20, 19-38.
- Fienberg, S.E., Johnson, M.S., et Junker, B.W. (1999). Classical multiple lists. *Journal of the Royal Statistical Society, A*, 162, 383-405.
- Frank, O., et Snijders, T.A.B. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.

Tableau 4
Biais relatif et racine carrée de l'erreur quadratique moyenne relative des estimateurs de variance

Population I											
$\bar{p}_1 \approx 0,05, \bar{p}_2 \approx 0,01$											
Taylor				Bootstrap				Taylor			
$\sqrt{r^2}$	rb	$\sqrt{r^2}$	rb	$\sqrt{r^2}$	rb	$\sqrt{r^2}$	rb	$\sqrt{r^2}$	rb	$\sqrt{r^2}$	rb
τ_1^r	NC	NC	0,01	0,17	NC	NC	-0,04	0,08	NC	NC	-0,20
τ_2^r	NC	NC	0,01	0,49	NC	NC	1,9 ^a	5,3 ^a	NC	NC	-0,02
τ_3^r	NC	NC	0,01	0,48	NC	NC	1,9 ^a	5,3 ^a	NC	NC	-0,02
τ_4^r	NC	NC	0,01	0,48	L ₁	L ₂	1,9 ^a	5,3 ^a	0,20	1,10	-0,02
τ_5^r	0,02	0,20	-0,01	0,17	0,03	0,19	-0,01	0,17	0,14	0,51	-0,06
τ_6^r	0,13	0,62	-0,01	0,49	0,24	1,20	1,7 ^a	4,6 ^a	0,22	0,92	-0,00
τ_7^r	0,13	0,61	-0,01	0,48	0,24	1,20	1,6 ^a	4,5 ^a	0,23	0,91	0,01
τ_8^r	0,06	0,21	0,02	0,17	0,05	0,19	-0,01	0,17	0,12	0,50	-0,08
τ_9^r	0,07	0,51	-0,03	0,44	-0,25	0,66	-0,11	1,40	0,13	0,69	-0,03
τ_{10}^r	0,06	0,50	-0,03	0,43	-0,25	0,66	-0,12	1,40	0,12	0,68	-0,03
τ_{11}^r	0,03	0,20	-0,01	0,17	0,03	0,18	-0,02	0,17	0,16	0,52	-0,05
τ_{12}^r	0,07	0,34	-0,02	0,35	-0,07	0,16	-0,03	0,12	0,10	0,42	-0,01
τ_{13}^r	0,06	0,34	-0,02	0,34	-0,05	0,14	-0,02	0,11	0,10	0,42	-0,01
τ_{14}^r	0,16	0,41	-0,03	0,41	-0,06	0,28	0,05	0,37	0,01	0,17	-0,01
τ_{15}^r	0,16	0,41	-0,03	0,41	-0,03	0,41	-0,03	0,15	0,01	0,16	0,16

Nota : rb, biais relatif, r^2 , erreur quadratique moyenne relative. Les indices supérieurs M et D des EMV τ_1, τ_2 , et τ indiquent des estimateurs de variance fondés sur le modèle et fondés sur le plan, respectivement. Intervalles de confiance bootstrap calculés sur 2 000 échantillons bootstrap, NC, non calculé. Résultats basés sur 10⁴ essais. L'indice supérieur a indique des résultats obtenus en ne tenant pas compte de 8 % des essais. Les essais omis étaient ceux pour lesquels l'estimateur correspondant de τ_2 était négatif ou supérieur à 10⁴. L₁ et L₂ indiquent des valeurs supérieures à 10² et 10⁴, respectivement.

Remerciements

Cette étude a été financée par la subvention UASIN-EXB-01-01 du PROMEP et par la subvention PAFI-UAS-2002-I-MHFM-0 de l'UAS. Nous remercions Eduardo Gutierrez, le rédacteur associé et les examinateurs de leurs suggestions et commentaires constructifs.

Bibliographie

Agresti, A. (2002). *Categorical Data Analysis*. 2^{ème} édition. New York : John Wiley & Sons, Inc.

Booth, J.G., Butler, R.W. et Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282-1289.

Castledine, B.J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 67, 197-210.

Daroch, J.N. (1958). The multiple-recapture census I: Estimation of a closed population. *Biometrika*, 45, 343-359.

Davison, A.C., et Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*. New York : Cambridge University Press.

Enfin, les meilleures propriétés de l'ensemble d'estimateurs τ_k^r sont une conséquence de la plus grande quantité d'information qu'ils utilisent. Bien que nous soyons servis de lois a priori relativement uniformes pour les τ_k , l'information qu'ils fournissent est suffisante pour éviter les problèmes de biais et de forte variabilité observés pour les autres estimateurs. Nous avons réalisé certains essais par simulations supplémentaires et les résultats (qui ne sont pas présentés dans les tableaux) indiquent qu'à condition que les lois a priori soient maintenues relativement uniformes, les estimations ne sont pas affectées par les valeurs de leurs paramètres. Manifestement, une information initiale erronée combinée à de faibles valeurs des $p_i^{(k)}$ aura une incidence sur les estimations. À titre d'exemple, mentionnons une loi a priori de τ_2 dont la densité de probabilité est fortement concentrée autour d'une valeur très éloignée de la valeur réelle de τ_2 . Cependant, nous pensons que si le chercheur dispose d'information correcte, même si elle est vague, il vaut la peine d'utiliser l'ensemble d'estimateurs τ_k^r .

Tableau 2 Biases relatif et racine carrée de l'erreur quadratique moyenne relative des estimateurs des tailles de population

Population I			Population II			Population III			Population IV		
\bar{p}_1	$\sqrt{rc_2}$	rb	$\sqrt{rc_2}$	rb	$\sqrt{rc_2}$	$\sqrt{rc_2}$	rb	$\sqrt{rc_2}$	$\sqrt{rc_2}$	rb	$\sqrt{rc_2}$
0,05	0,01	0,002	0,05	0,01	0,002	0,05	0,01	0,006	0,05	0,01	0,006

\bar{p}_1	-0,00	0,02	-0,00	0,06	-0,00	0,02	-0,00	0,06	-0,00	0,02	-0,01	0,09
\bar{r}_1	0,01	0,12	0,24 ^a	0,78 ^a	0,01	0,13	0,21 ^a	0,76 ^a	0,00	0,06	0,17 ^b	0,63 ^c
\bar{r}_2	0,01	0,07	0,13 ^a	0,43 ^a	0,01	0,07	0,12 ^a	0,42 ^a	0,00	0,02	0,05 ^b	0,18 ^c
\bar{r}_3	-0,00	0,02	-0,00	0,06	-0,00	0,02	-0,00	0,06	-0,00	0,02	-0,00	0,09
\bar{r}_4	0,02	0,13	0,14 ^a	0,65 ^a	0,01	0,12	0,14 ^a	0,65 ^a	0,00	0,06	0,13	0,71
\bar{r}_5	0,01	0,07	0,08 ^a	0,36 ^a	0,01	0,07	0,08 ^a	0,36 ^a	0,00	0,02	0,03	0,20
\bar{r}_6	-0,00	0,02	-0,01	0,06	-0,00	0,02	-0,01	0,06	-0,00	0,02	-0,01	0,09
\bar{r}_7	-0,00	0,02	-0,01	0,06	-0,00	0,02	-0,01	0,06	-0,00	0,02	-0,01	0,09
\bar{r}_8	-0,00	0,12	-0,14	0,48	-0,00	0,12	-0,14	0,48	-0,00	0,11	-0,04	0,35
\bar{r}_9	-0,00	0,07	-0,08	0,27	-0,00	0,07	-0,08	0,27	-0,00	0,02	-0,02	0,12
\bar{r}_{10}	-0,00	0,02	-0,01	0,06	-0,00	0,02	-0,01	0,06	-0,00	0,02	-0,01	0,09
\bar{r}_{11}	0,02	0,12	0,07	0,20	0,11	0,07	0,20	0,11	0,00	0,06	0,01	0,18
\bar{r}_{12}	0,02	0,07	0,20	0,11	0,07	0,20	0,11	0,00	0,00	0,06	0,00	0,18
\bar{r}_{13}	0,01	0,06	0,04	0,11	0,03	0,06	0,03	0,11	0,00	0,02	-0,00	0,08

Nota : rb, biais relatif; rc_2 , erreur quadratique moyenne relative; \bar{r}_1, \bar{r}_2 et \bar{r}_3 , EMV. Les indices supérieurs L_1, J et P des estimateurs \bar{r}_1, \bar{r}_2 et \bar{r}_3 indiquent des estimateurs bayésiens basés sur une loi uniforme, de Jeffreys et de Poisson-Gamma à deux degrés, respectivement. Les résultats sont fondés sur 10⁴ essais. Les indices supérieurs a, b et c indiquent des résultats obtenus en ne tenant pas compte de 8 %, de 15 % et de 21 % des essais. Les essais omis étaient ceux pour lesquels l'estimateur correspondant de \bar{r}_2 était négatif ou supérieur à 10⁴.

Tableau 3 Probabilité de couverture et longueur moyenne des intervalles de confiance à 95 %

Population I						Population II					
$\bar{p}_1 \approx 0,05, \bar{p}_2 \approx 0,01$			$\bar{p}_1 \approx 0,01, \bar{p}_2 \approx 0,002$			$\bar{p}_1 \approx 0,05, \bar{p}_2 \approx 0,01$			$\bar{p}_1 \approx 0,01, \bar{p}_2 \approx 0,002$		
Bootstrap	Wald	cp	Bootstrap	Wald	cp	Bootstrap	Wald	cp	Bootstrap	Wald	cp

\bar{r}_M	NC	NC	0,95	1,29	NC	NC	0,94	3,98	NC	NC	0,93	1,27	NC	NC	0,76	4,00
\bar{r}_M^2	NC	NC	0,95	1,044	NC	NC	0,90 ^a	8 181 ^a	NC	NC	0,95	1 029	NC	NC	0,90 ^a	7 764 ^a
\bar{r}_M^3	NC	NC	0,95	1 052	NC	NC	0,90 ^a	8 200 ^a	NC	NC	0,95	1 037	NC	NC	0,90 ^a	7 784 ^a
\bar{r}_M^4	0,95	1,30	0,95	1,29	0,92	3,99	0,94	4,04	0,97	1,47	0,95	1,37	0,96	6,42	0,92	6,57
\bar{r}_M^5	0,94	1 110	0,95	1 044	0,74	L_1	0,90 ^a	8 181 ^a	0,94	1 129	0,95	1 029	0,74	L_1	0,90 ^a	7 764 ^a
\bar{r}_M^6	0,94	1 118	0,95	1 052	0,75	L_1	0,90 ^a	8 201 ^a	0,95	1 139	0,95	1 038	0,78	L_1	0,90 ^a	7 819 ^a
\bar{r}_M^7	0,94	1 31	0,95	1,29	0,92	4,12	0,94	4,03	0,97	1,50	0,94	1,37	0,97	6,68	0,93	6,57
\bar{r}_M^8	0,94	1 116	0,95	1 049	0,72	L_2	0,89 ^a	6 887 ^a	0,94	1 128	0,95	1 028	0,73	L_2	0,89 ^a	6 738 ^a
\bar{r}_M^9	0,94	1 124	0,95	1 057	0,73	L_2	0,90 ^a	6 908 ^a	0,95	1 139	0,95	1 038	0,77	L_2	0,90 ^a	6 796 ^a
\bar{r}_M^{10}	0,95	1,31	0,95	1,28	0,93	4,12	0,94	4,02	0,96	1,51	0,95	1,37	0,96	6,66	0,92	6,52
\bar{r}_M^{11}	0,93	1 043	0,94	998	0,58	3 122	0,71	3 142	0,93	1 057	0,93	985	0,60	3 074	0,72	3 095
\bar{r}_M^{12}	0,93	1 052	0,94	1 007	0,60	3 199	0,72	3 178	0,94	1 072	0,93	995	0,68	3 276	0,73	3 188
\bar{r}_M^{13}	0,94	1 31	0,95	1,29	0,91	4,11	0,94	4,02	0,89	1,51	0,95	1,37	0,86	6,66	0,93	6,54
\bar{r}_M^{14}	0,97	997	0,95	957	1,00	1 506	0,92	1 573	0,97	1 000	0,95	943	1,00	1 510	0,92	1 577
\bar{r}_M^{15}	0,97	1 006	0,95	966	1,00	1 575	0,94	1 624	0,97	1 011	0,95	953	1,00	1 679	0,95	1 710

Nota : cp, probabilité de couverture; \bar{r}_1 , longueur moyenne. Les indices supérieurs M et D des EMV \bar{r}_1, \bar{r}_2 et \bar{r}_3 indiquent des intervalles de confiance fondés sur le modèle et fondés sur le plan, respectivement. Intervalles de confiance bootstrap calculés sur 2 000 échantillons bootstrap, NC, non calculé. Résultats fondés sur 10⁴ essais. L'indice supérieur a indique des résultats obtenus en ne tenant pas compte de 8 % des essais. Les essais omis sont ceux pour lesquels l'estimateur correspondant de \bar{r}_2 était négatif ou supérieur à 10⁴. L_1 et L_2 indiquent des longueurs supérieures à 10⁴ et 10⁵, respectivement.

rédacteur associé, ce modèle implique que le nombre de personnes nommées par la grappe J_i est l'espérance $(1 - m_i) (1 - \exp(-\beta^1 m_i)) + \tau_2 (1 - \exp(-\beta^2 m_i))$ et que, par conséquent, le nombre de personnes nommées est approximativement proportionnel à m_i . Notons que le modèle échangeable hypothétique pour $p_{(k)}^i$ ne postule pas ce genre de relation avec m_i . Puisque l'estimation de $p_{(k)}^i$ dépend principalement de $z_{(k)}^i$, le nombre de personnes dans U_k nommées par la grappe J_i , nous nous attendons à ce que l'omission de cette relation n'ait pas d'incidence sur l'efficacité de l'estimateur de $p_{(k)}^i$. Darroch (1958) a montré, dans le cas de l'estimation du maximum de vraisemblance, que l'on ne réalise aucun gain significatif en émettant l'hypothèse d'un modèle prises-effort.

Tableau I

Paramètres des populations simulées		Population III Population IV		Population I		Population II	
$N = 250$	$N = 250$	$N = 250$	$N = 250$	$N = 250$	$N = 250$	$N = 250$	$N = 250$
M_i , Poisson	M_i , Poisson	$E(M_i) = 7,2$	$E(M_i) = 7,2$	$E(M_i) = 7,2$	$E(M_i) = 7,2$	$V(M_i) = 7,2$	$V(M_i) = 7,2$
M_i , Binomiale nég.	M_i , Binomiale nég.	$V(M_i) = 24,48$	$V(M_i) = 24,48$	$V(M_i) = 24,48$	$V(M_i) = 24,48$	$\tau_1 = 1,811$	$\tau_1 = 1,811$
		$\tau_2 = 2,200$	$\tau_2 = 2,200$	$\tau_2 = 2,200$	$\tau_2 = 2,200$	$\tau_1 / \tau = 0,45$	$\tau_1 / \tau = 0,46$
		$\tau = 2,511$	$\tau = 2,511$	$\tau = 2,511$	$\tau = 2,511$		
		$\tau_1 / \tau = 0,72$	$\tau_1 / \tau = 0,72$	$\tau_1 / \tau = 0,72$	$\tau_1 / \tau = 0,72$		

Pour les populations I et II, les valeurs des paramètres des lois a priori étaient $\sigma_k^2 = 2,5$, $\mu_k = -3,5$, $\gamma_k^2 = 25$, $k = 1, 2$, $a_1 = 1$, $b_1 = 0,1$, $a_2 = 7,84$, $b_2 = 0,0082$, de sorte que $E(\lambda_1) = 10$, $V(\lambda_1) = 100$, $E(\lambda_2) = 2,800$, et $V(\lambda_2) = 10^6$. Pour les populations III et IV les valeurs des paramètres étaient $\sigma_k^2 = 9$, $\mu_k = -3,5$, $\gamma_k^2 = 9$, $k = 1, 2$, $a_1 = 1$, $b_1 = 0,1$, $a_2 = 8$, $b_2 = 0,01$, de sorte que $E(\lambda_1) = 10$, $V(\lambda_1) = 100$, $E(\lambda_2) = 800$ et $V(\lambda_2) = 80\,000$. Ces valeurs impliquent que les lois a priori sont bien dispersées sur les intervalles relativement grands qui contiennent les paramètres d'intérêt.

Nous avons réalisé l'expérience par simulation comme il suit. À partir de chaque population de $N = 250$ valeurs de m_i , nous avons sélectionné un EASSR de $n = 25$ valeurs. À partir de la grappe J_i dans l'échantillon, nous avons généré les valeurs de $X_{(k)}^{(i)}$ et $X_{(k)}^{(j)}$ en tirant des échantillons de taille $\tau_1 - m_i$ et τ_2 à partir de lois de Bernoulli de moyenne $p_{(k)}^{(i)}$ et $p_{(k)}^{(j)}$, respectivement. Ces données ont été utilisées pour calculer les estimateurs suivants des tailles de population : l'ensemble d'EMV $\hat{\tau}_1, \hat{\tau}_2$, et $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$ proposé par Félix-Médina et Thompson (2004); ainsi que les trois ensembles d'estimateurs bayésiens $\hat{\tau}_1^b, \hat{\tau}_2^b$ et $\hat{\tau}^b = \hat{\tau}_1^b + \hat{\tau}_2^b$, $a = U, J, P$, obtenus en utilisant comme lois a priori les lois uniforme (U), de Jeffreys (J) et de Poisson (P) respectivement. En outre, nous avons calculé

les estimateurs de variance et les intervalles de confiance. Nous avons calculé les intervalles bootstrap selon la méthode de base, sauf les intervalles fondés sur les estimateurs $\hat{\tau}_1^b, \hat{\tau}_2^b$ et $\hat{\tau}^b$, qui ont été calculés par la méthode des centiles. Tous les estimateurs bootstrap ont été obtenus en utilisant 2 000 échantillons bootstrap. Enfin, les propriétés des estimateurs ponctuels et d'intervalle ont été évaluées en utilisant $r = 10\,000$ essais de la méthode qui précède.

Nous avons évalué les propriétés d'un estimateur, disons $\hat{\tau}$, par son biais relatif et la racine carrée de son erreur quadratique moyenne relative, définis comme étant $\sqrt{r - \text{eqm}} = \sqrt{r - \text{eqm}}$ et $\sqrt{\sum_{i=1}^r (\hat{\tau}_i - \tau)^2 / (r - 1)}$, où $\hat{\tau}_i$ est la valeur de $\hat{\tau}$ obtenue lors du i^{e} essai. Nous avons également évalué les propriétés d'un estimateur de variance par son biais relatif et la racine carrée de son erreur quadratique moyenne relative, qui ont été définis de la même façon que ceux d'un estimateur de la taille de population, mais en utilisant la variance déterminée empiriquement plutôt que la variance réelle. Enfin, nous avons évalué les propriétés des intervalles de confiance à 95 % par leur probabilité de couverture et leur longueur moyenne.

6. Résultats et discussion

Faute d'espace, aux tableaux 2 à 4, nous présentons que certains résultats de l'étude numérique. Toutefois, les commentaires qui suivent ont trait à l'ensemble complet de résultats.

Malgré les limites de l'étude par simulation, nous pouvons conclure que le principal facteur qui influe sur les propriétés des estimateurs et des intervalles de confiance et la grandeur des $p_{(k)}^i$. Lorsque celles-ci sont grandes et indépendamment de la loi des M_i et de la taille de la fraction τ_1 / τ couverte par la base de sondage, chacun des estimateurs des τ et des intervalles de confiance de type fondé sur le plan de sondage (Wald ou bootstrap) donne de bons résultats. Toutefois, lorsque les $p_{(k)}^i$ sont faibles et malgré tous les autres facteurs, seuls les estimateurs bayésiens $\hat{\tau}_k^b$ donnent des résultats acceptables. Il mérite d'être souligné que, si les $p_{(k)}^i$ sont faibles, les estimateurs bayésiens $\hat{\tau}_k^b$ et $\hat{\tau}^b$ donnent de meilleurs résultats que l'EMV $\hat{\tau}_k$; toutefois, les propriétés de $\hat{\tau}_k^b$ et $\hat{\tau}^b$ ne sont pas suffisamment bonnes pour rendre les inférences fiables. Les intervalles de confiance bootstrap pour τ_1 basés sur $\hat{\tau}_1^b$ ne sont pas aussi bons que les intervalles de Wald lorsque les $p_{(k)}^i$ sont faibles ou que les M_i ne suivent pas une loi de Poisson. L'explication de ce résultat et l'établissement de meilleurs intervalles bootstrap sont des sujets qui devront être étudiés de façon plus approfondie.

$$(10) \quad \left\{ \begin{aligned} & \sum_{i=1}^n \left(\frac{p_{(2)}^i}{1 - \bar{d}_z} - \frac{\bar{d}_z}{\bar{d}_z} \right) \left(\bar{t}_z p_{(2)}^i (1 - \bar{d}_z) \right) \\ & + \left[\frac{(\bar{t}_z - r_z)^2}{\bar{t}_z \bar{Q}_z (1 - \bar{Q}_z)} - \frac{2 \bar{t}_z \bar{Q}_z}{\bar{d}_z} \right] \left(\frac{\bar{d}_z}{\bar{d}_z} \right) \left(\bar{d}_z \right) \end{aligned} \right\}$$

où $\bar{Q}_z = \Pi_{i=1}^n p_{(2)}^i$,

$$\begin{aligned} \bar{A}_z &= \sum_{i=1}^n \frac{\bar{d}_z}{(p_{(2)}^i)^2} - \bar{C}_z + \frac{\bar{t}_z + a_z - 1}{1} - \frac{\bar{t}_z - r_z}{1} \\ \bar{C}_z &= \frac{(v_z^{-1} - n \sigma_z^2) \left[n \sum_{i=1}^n p_{(2)}^i / \bar{d}_z \right]}{1 + n \sum_{i=1}^n p_{(2)}^i / \bar{d}_z}, \\ \bar{d}_z &= \frac{n^{-1} (v_z^{-1} - n \sigma_z^2) \sum_{i=1}^n p_{(2)}^i / \bar{d}_z}{1 + n \sum_{i=1}^n p_{(2)}^i / \bar{d}_z}, \end{aligned}$$

et

$$(8) \quad \bar{C}_1 = \frac{(v_1^{-1} - n \sigma_1^2) \left[n \sum_{i=1}^n p_{(1)}^i / \bar{d}_1 \right]}{1 + n \sum_{i=1}^n p_{(1)}^i / \bar{d}_1}.$$

En outre, puisque $\text{Cov}(Z_{(1)}^i, R_1 | \mathbf{m}_s) = (r_1 - m) \bar{Q}_1 p_{(1)}^i$, un

estimateur de $E[p[V_\varepsilon(\bar{t}_1 | \mathbf{m}_s)]]$ a la forme

$$(9) \quad \left\{ \begin{aligned} & \sum_{i=1}^n \left(\frac{p_{(1)}^i}{\bar{d}_1} \right) \left(\bar{t}_1 - m_i \right) \bar{d}_1 (1 - \bar{d}_1) \\ & + \left[\frac{(\bar{t}_1 - m - m_i)^2}{(\bar{t}_1 - m) \bar{Q}_1 (1 - \bar{Q}_1)} - \frac{2(\bar{t}_1 - m) \bar{Q}_1}{\bar{d}_1} \right] \left(\frac{\bar{d}_1}{\bar{d}_1} \right) \left(\bar{d}_1 \right) \end{aligned} \right\}$$

où

$$\bar{d}_1 = \frac{n^{-1} (v_1^{-1} - n \sigma_1^2) \sum_{i=1}^n p_{(1)}^i / \bar{d}_1}{1 + n \sum_{i=1}^n p_{(1)}^i / \bar{d}_1}.$$

Par conséquent, un estimateur de type fondé sur le plan de

sondage de $V(\bar{t}_1)$ est $\hat{V}_{11} + \hat{V}_{12}$.

Dans le cas de \bar{t}_z^* , puisque $Z_{(2)}^i | \mathbf{m}_s \sim \text{bin}(\tau_z, p_{(2)}^i)$ et $R_2 | \mathbf{m}_s \sim \text{bin}(\tau_z, 1 - \bar{Q}_z)$, ou $\bar{Q}_z = \Pi_{i=1}^n (1 - p_{(2)}^i)$, il s'ensuit que $E_p[V_\varepsilon(\bar{t}_z^* | \mathbf{m}_s)]$ ne dépend pas de \mathbf{m}_s , et conséquemment que $V_p[E_\varepsilon(\bar{t}_z^* | \mathbf{m}_s)] \approx 0$. Donc, puisque $\text{Cov}(Z_{(2)}^i, R_2 | \mathbf{m}_s) = \tau_z \bar{Q}_z p_{(2)}^i$, un estimateur de $V(\bar{t}_z)$ est

Enfin, puisque la non-dépendance de $E_\varepsilon(\bar{t}_z^* | \mathbf{m}_s)$ par rapport à \mathbf{m}_s implique que $\text{Cov}(\bar{t}_1^*, \bar{t}_z^*) \approx 0$, il s'ensuit qu'un estimateur de la variance de \hat{t} est $\hat{V}(\hat{t}) = \hat{V}(\bar{t}_1) + \hat{V}(\bar{t}_z)$.

5. Étude de Monte Carlo

Nous avons considéré quatre populations, décrites chacune au tableau I. Dans la paire formée par les populations I et II, la base de sondage couvrait environ 45 % de la population, tandis que dans la paire formée par les populations III et IV, elle couvrait environ 70 % de la population. Les populations de chaque paire étaient fort semblables, sauf le fait que, dans l'une des populations de chaque paire, la loi des M_i était une loi de Poisson, tandis que dans l'autre il s'agissait d'une loi binomiale négative. Les probabilités de nomination $p_{(k)}^i, i = 1, \dots, N, k = 1, 2$, ont été générées en utilisant le modèle $p_{(k)}^i = 1 - \exp(-\beta_k m_i)$, où les valeurs de β_k étaient fixées de façon que les valeurs suivantes de $\bar{p}_{(k)}^i = \sum_{i=1}^N p_{(k)}^i / N$ soient obtenues. Pour les populations I et II : $(\bar{p}_{(1)}^i, \bar{p}_{(2)}^i) \approx (0, 05, 0, 01)$ et $(\bar{p}_{(1)}^i, \bar{p}_{(2)}^i) \approx (0, 01, 0, 002)$. Pour les populations III et IV : $(\bar{p}_{(1)}^i, \bar{p}_{(2)}^i) \approx (0, 05, 0, 03)$ et $(\bar{p}_{(1)}^i, \bar{p}_{(2)}^i) \approx (0, 01, 0, 006)$. Le modèle employé pour générer les $p_{(k)}^i$ est un modèle utilisé dans les méthodes prises-effort (voir Seber 1982, chapitre 7 pour une description de ces méthodes). Comme l'a souligné un

où $\hat{\alpha}_{(k)}^1 = \sum_{i=1}^n \alpha_{(k)}^1 / m, k = 1, 2$. Il s'ensuit qu'un estimateur de τ est $\hat{\tau} = \hat{\tau}_1 + \hat{\tau}_2$.
Les formes de ces estimateurs sont fondamentalement des ajustements des formes de l'EMV proposées par Félix-Medina et Thompson (2004), de sorte qu'est intégrée dans les estimateurs proposés l'information initiale au sujet de τ_k et $\alpha_{(k)}^1, i = 1, \dots, n, k = 1, 2$. En outre, comme la fait remarquer un examinateur, l'estimateur $\hat{p}_{(k)}^1$ a la forme de l'EMV de $p_{(k)}^1$ suivi par des termes de rétroécassement, l'un étant celui de $\alpha_{(k)}^1$ vers la moyenne arithmétique $\hat{\alpha}_{(k)}^1$ et l'autre, celui de $\hat{\alpha}_{(k)}^1$ vers la moyenne a priori μ_k .

4. Intervalles de confiance pour les tailles de population

Comme nous l'avons indiqué plus haut, nous utiliserons l'approche fréquentiste pour obtenir des intervalles de confiance de type fondé sur le plan de sondage qui sont robustes aux écarts par rapport à la loi de Poisson hypothétique des M_i . Nous examinerons des intervalles de confiance et des intervalles de Wald basés sur une approximation normale (voir Agresti 2002, page 13 et Evans, Kim et O'Brien 1996 pour la terminologie la plus récente).

4.1 Intervalles de confiance bootstrap

Nous utiliserons une version du bootstrap obtenue en combinant la variante du bootstrap pour les populations finies proposée par Booth, Butler et Hall (1994) et la variante paramétrique du bootstrap (voir Davison et Hinkley

1997, chapitre 2).
Les étapes de la méthode que nous proposons sont les suivantes. (i) Construire une population artificielle de N valeurs des m_i en répétant N/n fois, en supposant que N/n est un nombre entier, l'échantillon sélectionné de n tailles de grappe m_1, \dots, m_n . Si $N = kn + r$, où k et r sont des entiers positifs, construire la population en répétant k fois l'échantillon sélectionné de n tailles de grappe et ajouter à cet ensemble de tailles m_i un échantillon aléatoire simple sans remise (EASSR) de r valeurs des m_i sélectionnées à partir de l'échantillon observé de n tailles de grappe. (ii) Sélectionner un EASSR de n tailles à partir de la population des m_i . Soit i_1, \dots, i_n les indices des m_i dans l'échantillon. (iii) Pour chaque $i = i_1, \dots, i_n$, tirer des échantillons de tailles $\hat{\tau}_1 - m_i$ et $\hat{\tau}_2$ à partir des lois de Bernoulli de moyennes $p_{(1)}^{(i)}$ et $p_{(2)}^{(i)}$, respectivement, où $\hat{\tau}_1, \hat{\tau}_2, p_{(1)}^{(i)}$ et $p_{(2)}^{(i)}$ sont les estimations de $\tau_1, \tau_2, p_{(1)}^{(i)}$ et $p_{(2)}^{(i)}$ calculées d'après l'échantillon observé original. Ces échantillons simulent les valeurs des ensembles $\{x_{ij}^{(1)}\}$ et $\{x_{ij}^{(2)}\}$ de variables indicatrices. (iv) Calculer les estimations de τ_1, τ_2 et τ à partir des échantillons tirés aux étapes (iii) et (iv) en suivant la même méthode que celle utilisée pour

calculer les estimations originales $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$. (v) Obtenir les lois bootstrap de $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$ en répétant les étapes (i) à (iv) un grand nombre B de fois et en calculant les lois empiriques à partir des ensembles de B valeurs de $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$. (vi) Construire les intervalles de confiance bootstrap de $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$ en utilisant la méthode de $100(1 - \alpha)\%$ pour τ_1, τ_2 et τ en utilisant la méthode de base ou celle des centiles (voir Davison et Hinkley 1997, chapitre 5 pour la description de ces méthodes). Dans la méthode de base, l'intervalle pour τ est $[\hat{\tau} - \hat{\tau}_{(1-\alpha/2)}, \hat{\tau} + \hat{\tau}_{(1-\alpha/2)}]$, et dans la méthode des centiles, il est $[\hat{\tau}_{(\alpha/2)}, \hat{\tau}_{(1-\alpha/2)}]$, où $\hat{\tau}_{(\alpha/2)}$ et $\hat{\tau}_{(1-\alpha/2)}$ sont les points $\alpha/2$ inférieur et supérieur de la distribution bootstrap de l'estimation originale $\hat{\tau}$ de τ .
Notons que cette variante du bootstrap ne repose pas sur l'utilisation de la loi de Poisson hypothétique des M_i , mais sur le plan d'échantillonnage utilisé pour sélectionner l'échantillon initial de grappes. Donc, nous pouvons conclure que les intervalles de confiance résultants sont robustes aux écarts par rapport à la loi hypothétique des M_i .
Si l'on souhaite également calculer les estimations bootstrap des variances de $\hat{\tau}_1, \hat{\tau}_2$ et $\hat{\tau}$, il est possible d'obtenir des estimations simples en calculant les variances d'échantillon des ensembles de B valeurs de ces estimateurs.

4.2 Intervalles de confiance de Wald

Bien que nous ne démontrions pas théoriquement ici que les estimateurs proposés des tailles de population suivent asymptotiquement une loi normale, nous supposons que la loi normale est une approximation raisonnable pour les tailles de population des intervalles de confiance de Wald $100(1 - \alpha)\%$ de type fondé sur le plan de sondage de la forme $\hat{\tau}_k \pm z_{1-\alpha/2} \sqrt{V(\hat{\tau}_k)}$, où $z_{1-\alpha/2}$ est le point $\alpha/2$ supérieur de la loi normale standard, et $V(\hat{\tau}_k)$ est un estimateur de type fondé sur le plan de sondage de la variance de $\hat{\tau}_k$.

Pour construire ce genre d'intervalle, nous commençons par dériver des estimateurs de la variance de type fondé sur le plan de sondage en suivant la même stratégie que celle utilisée par Félix-Medina et Thompson (2004). Celle-ci consiste à remplacer la distribution des tailles de grappes par celle du plan d'échantillonnage utilisé pour sélectionner l'échantillon initial S_0 . Nous utilisons pour cela la formule :

$$V(\hat{\tau}_k) = V_p[E_{\hat{\tau}_k}(\hat{\tau}_k | m_s)] + E_p[V_p(\hat{\tau}_k | m_s)], \quad (6)$$

où $E_{\hat{\tau}_k}(\hat{\tau}_k | m_s)$ et $V_{\hat{\tau}_k}(\hat{\tau}_k | m_s)$ dénotent les opérateurs d'espérance et de variance conditionnelles fondées sur le modèle, sachant que $M_s = m_s$; et $E_p(\cdot)$ et $V_p(\cdot)$ dénotent les opérateurs d'espérance et de variance fondés sur le plan de sondage. Donc, nous obtenons les estimateurs de variance en appliquant (6) aux approximations de Taylor de

$\pi(\tau_k) \propto 1$, où $k=1, 2$, et τ_1 et τ_2 sont des variables

aléatoires indépendantes.

La loi a priori de Poisson de r_1 détermine dans le premier cas la pour motivation le fait que $r_1 = \sum_{i=1}^N M_i$, et que M_i est une variable de Poisson de moyenne λ_i . Souignons que ce cas permet aux chercheurs d'utiliser l'information au sujet de r_1 et r_2 qui est connue avant l'observation de l'échantillon. Par ailleurs, les lois définies dans les deux autres cas ne sont pas informatives.

Dans le cas des probabilités de nomination $D_{(i)}^{(k)}$, à l'instar de Castledine (1981), nous supposons qu'elles sont échangeables et nous utilisons le modèle normal à deux degrés pour les logits $\alpha_{(i)}^{(k)} = \log [D_{(i)}^{(k)} / (1 - D_{(i)}^{(k)})]$ des $D_{(i)}^{(k)}$:

$$\alpha_{(i)}^{(k)} | \theta \sim N(\theta) \quad \theta \sim \mathcal{O}_2^{(k)}$$

$$(1) \quad \theta^k \sim N(\mu^k, \gamma^k \Sigma^k); i = 1, \dots, n, k = 1, 2,$$

(1) en fixant $\theta^k = \mu^k$ et $\gamma^k_z = 0$, $k = 1, 2$.

Enfin, nous supposons que tous les vecteurs aléatoires (τ_k, λ_k) et (α_k, θ_k) , où $\alpha_k = (\alpha_{(k)}^1, \dots, \alpha_{(k)}^n)$, $k = 1, 2$, sont mutuellement indépendants.

Bien que nous ayons défini trois types de loi a priori pour τ_1 et τ_2 , elles peuvent être traitées de manière uniforme, parce que les lois marginales a priori de τ_1 et τ_2 obtenues à partir des lois de Poisson-Gamma, sont les lois binomiales

negatives :

$$i_2 \left(\frac{i_1 q + N}{N} \right) \frac{i_1 i_2}{(i_1 q + i_1) \mathbf{I}} \propto (i_1) u$$

et

$$(7) \quad \left(\frac{1}{1 + q^{z_1}} \right) \frac{q^{z_1}}{\Gamma(\tau_2 + a_2)} \propto (\tau_2) u$$

où $\Gamma(\cdot)$ dénote la fonction Gamma. Les lois de Jeffreys et les lois uniformes sont des cas limites de (2) obtenus en supposant que $a_k = b_k = 0, k = 1, 2$, et $a_k = 1, b_k = 0$,

$k=1, 2$, respectivement. Notons que la loi Gamma n'est pas définie pour ces valeurs de a_k et b_k ; toutefois, pour la dérivation des estimateurs, nous pouvons utiliser ces valeurs

exprimée sous la forme

$$\begin{aligned} & \left[\frac{z^{\iota_2} \mathcal{Z}}{z^{(\iota_2)} \underline{\mathfrak{z}} - {}^{(\iota_2)} \mathfrak{z}} \right] \frac{\mathrm{d} x \mathfrak{e} \frac{\left[{}^{(\iota_2)} z^{\iota_1} \mathfrak{a} \right] \mathrm{d} x \mathfrak{e} + \mathbb{I}}{z^{(\iota_2)} \underline{\mathfrak{z}} - {}^{(\iota_2)} \mathfrak{z}}}{\sum_u {}^{(\iota_2)} z^{\iota_1} \mathfrak{a}} \prod_u {}^{(\iota_2)} z^{\iota_1} \mathfrak{a} \\ & \frac{z^{\iota_2} \mathcal{Z}}{z^{(\iota_2)} \underline{\mathfrak{z}} - {}^{(\iota_2)} \mathfrak{z}} \frac{\mathrm{d} x \mathfrak{e} \frac{\left[{}^{(\iota_2)} z^{\iota_1} \mathfrak{a} \right] \mathrm{d} x \mathfrak{e} + \mathbb{I}}{z^{(\iota_2)} \underline{\mathfrak{z}} - {}^{(\iota_2)} \mathfrak{z}}}{\sum_u {}^{(\iota_2)} z^{\iota_1} \mathfrak{a}} \prod_u {}^{(\iota_2)} z^{\iota_1} \mathfrak{a} \\ & \frac{z^{\iota_2} \mathcal{Z}}{z^{(\iota_2)} \underline{\mathfrak{z}} - {}^{(\iota_2)} \mathfrak{z}} \frac{\mathrm{d} x \mathfrak{e} \frac{\left[{}^{(\iota_2)} z^{\iota_1} \mathfrak{a} \right] \mathrm{d} x \mathfrak{e} + \mathbb{I}}{z^{(\iota_2)} \underline{\mathfrak{z}} - {}^{(\iota_2)} \mathfrak{z}}}{\sum_u {}^{(\iota_2)} z^{\iota_1} \mathfrak{a}} \prod_u {}^{(\iota_2)} z^{\iota_1} \mathfrak{a} \end{aligned}$$

$$\frac{\frac{1}{\epsilon} \mathcal{D}({}^i W - \frac{1}{2})}{({}_0 \frac{\mathcal{D}}{\epsilon}) - ({}_0 \frac{1}{\epsilon} \mathcal{D})} - \frac{{}^i W - \frac{1}{2}}{({}_0 \frac{1}{\epsilon} Z)} = \frac{\{({}_0 \frac{1}{\epsilon} \mathcal{D})\} d\mathbf{x} \otimes \mathbf{I}}{\{({}_0 \frac{1}{\epsilon} \mathcal{D})\} d\mathbf{x} \otimes \mathbf{I}} = ({}_0 \frac{1}{\epsilon} d$$

$$\begin{aligned}
\frac{\frac{\zeta_2 \mathcal{O} \zeta_2}{(\zeta_2) \mathcal{O} - (\zeta_2)' \mathcal{O}} - \frac{\zeta_2}{(\zeta_2)' \mathcal{Z}}}{\frac{(\zeta_2)' \mathcal{D} - 1}{(\zeta_2)' \mathcal{D} - 1} \prod_{u=1}^{l-1} [(\zeta q + 1) / 1] - 1} &= \frac{\zeta_2}{(\zeta_2)' \mathcal{D} - 1} \prod_{u=1}^{l-1} [(\zeta q + 1) / (1 - \zeta v_u)] + \zeta_R \\
\zeta_u, \dots, 1 = 1 : \frac{l \lambda ({}^l \mathcal{W} - \zeta_2) u}{1! n - (1) \mathcal{O}} &
\end{aligned}$$

(5)

2. Plan d'échantillonnage et notation

La structure de la population et le plan d'échantillonnage que nous considérons dans le présent article sont les mêmes que ceux proposés par Félix-Medina et Thompson (2004), où nous avons une brève description. Soit $U = \{u_1, \dots, u_n\}$ une population humaine cachée de taille inconnue n . Soit U_1 un sous-ensemble de U formé par un nombre inconnu r_1 de personnes que l'on peut trouver dans différents emplacements accessibles, comme des bars, des parcs ou des îlots d'habitations. Ce plan d'échantillonnage s'appuie sur les hypothèses qu'il est possible de construire une base de sondage de N de ces emplacements et que le chercheur établit une règle opérationnelle qui lui permet de déterminer si une personne appartient ou non à un emplacement figurant dans la base de sondage et, dans l'affirmative, de situer cet emplacement. Souignons que l'on ne suppose pas que le sous-ensemble U_1 couvert par la base de sondage représente la partie principale de U et que, comme dans l'échantillonnage en grappes ordinaire, on suppose qu'une personne figurant dans la base de sondage n'appartient qu'à un seul emplacement. Soit A_i le i -ième emplacement ou grappe dans la base de sondage et m_i le nombre de personnes qui appartiennent à A_i , $i = 1, \dots, N$; alors $r_1 = \sum_{i=1}^N m_i$. Enfin, soit $U_2 = U - U_1$ la partie de U non couverte par la base de sondage et soit $r_2 = n - r_1$ sa taille.

Le plan d'échantillonnage est le suivant. Un échantillon $S_0 = \{A_1, \dots, A_n\}$ de n grappes est tiré à partir de la base de sondage par échantillonnage aléatoire simple sans remise, et les m_i personnes qui appartiennent à chaque $A_i \in S_0$ sont identifiées. Notons que nous avons utilisé les indices $1, \dots, n$ pour dénoter les grappes dans S_0 ; toutefois, cela ne signifie pas que les n premières grappes dans la base de sondage sont nécessairement les grappes contenues dans l'échantillon. Puis, on demande aux personnes membres dans la grappe échantillonnée A_i de nommer des membres de U , mais seules les personnes nommées comprises dans $U - A_i$ sont prises en considération. Cette procédure est répétée pour chaque grappe $A_i \in S_0$. Par convention, nous dirons qu'une personne est nommée par une grappe si elle est nommée par au moins un membre de cette grappe. Les nominations à partir des diverses grappes sont faites indépendamment les unes des autres, et diverses stratégies de nomination peuvent être utilisées à différents emplacements. Par exemple, à l'emplacement A_i , les nominations pourraient être faites par les m_i membres, en tant que groupe, tandis que dans un autre emplacement, A_j , chacun des m_j membres pourrait faire des nominations séparément. Enfin, pour chaque personne nommée, le chercheur doit enregistrer le ou les emplacements qui l'ont nommée, ainsi que le segment U_1 ou U_2 de la population auquel elle appartient. Il convient de souligner que ce dernier élément

Un échantillon aléatoire simple d'emplacements est sélectionné et les membres de la population appartenant à chaque emplacement sont identifiés. Enfin, comme dans l'échantillonnage par dépiçage de liens ordinaire, il est demandé aux personnes se trouvant à chaque emplacement de nommer d'autres membres de la population.

Ces auteurs ont dérivé des estimateurs du maximum de vraisemblance (EMV) des tailles de population à partir de modèles probabilistes qui décrivent le nombre d'éléments découverts à chaque emplacement, ainsi que la probabilité qu'un membre de la population soit nommé à un emplacement, à laquelle on donne le nom de probabilité de nomination. Ils proposent aussi des estimateurs de variance fondés sur un modèle et partiellement fondés sur le plan de sondage, c'est-à-dire des estimateurs fondés à la fois sur le plan de sondage utilisé pour sélectionner l'échantillon initial et sur les modèles hypothétiques. Tout au long de l'article, nous parlerons d'estimateur « de type fondé sur le plan » pour faire référence à ce genre d'estimateur. Grâce à une étude par simulation, les auteurs ont montré que les EMV des tailles de population et leurs estimateurs de variance de type fondé sur le plan sont robustes aux écarts par rapport au modèle hypothétique, mais que les estimateurs de variance fondés sur un modèle ne sont pas robustes. En outre, ils ont constaté que les EMV ont tendance à surestimer gravement la taille de population si les probabilités de nomination sont faibles.

Comme l'ont indiqué ces auteurs, le problème de la surestimation qui se manifeste lorsque les probabilités de nomination sont faibles est dû à la petite quantité d'information que contient l'échantillon, quantité qui n'est pas suffisante pour obtenir des estimations stables des probabilités de nomination. Selon eux, un remède éventuel à ce problème consiste à suivre l'approche bayésienne pour construire des estimateurs dans lesquels sont intégrés des renseignements supplémentaires au sujet des paramètres de la population.

Ici, nous utilisons la méthode bayésienne pour faciliter la construction des estimateurs des tailles de population, mais nous faisons les inférences selon une approche fréquentiste. Donc, en plus de calculer des estimateurs ponctuels, nous construisons des intervalles de confiance. Nous adoptons pour cela la stratégie proposée par Félix-Medina et Thompson (2004) en vue de construire des intervalles de confiance basés sur la loi normale et utilisons des estimateurs de variance de type fondé sur le plan obtenus par la méthode de confiance bootstrap de type fondé sur le plan. Nous disons que cette approche inférentielle est « assistée par la méthode bayésienne ».

Combinaison de l'échantillonnage par dépistage de liens et de cachées : Une approche assistée par la méthode bayésienne

Martín H. Félix-Medina et Pedro E. Monjardín¹

Résumé

Félix-Medina et Thompson (2004) ont proposé une variante de l'échantillonnage par dépistage de liens dans laquelle on suppose qu'un part de la population (qui n'est pas nécessairement la plus grande) est couverte par une liste d'emplacements disponibles où les membres de la population peuvent être trouvés avec une probabilité élevée. Après la sélection d'un échantillon d'emplacements, on demande aux personnes se trouvant à chacun de ces emplacements de nommer d'autres membres de la population. Les deux auteurs ont proposé des estimateurs du maximum de vraisemblance des tailles de population qui donnent des résultats acceptables à condition que, pour chaque emplacement, la probabilité qu'un membre de la population soit nommé par une personne se trouvant à cet emplacement, appelée probabilité de nomination, ne soit pas faible. Dans la présente étude, nous partons de la variante de Félix-Medina et Thompson, et nous proposons trois ensembles d'estimateurs des tailles de population dérivés sous une approche bayésienne. Deux des ensembles d'estimateurs sont obtenus en utilisant des lois a priori incorrectes des tailles de population, et l'autre en utilisant des lois a priori de Poisson. Cependant, nous n'utilisons la méthode bayésienne que pour faciliter la construction des estimateurs et adoptons l'approche fréquentiste pour faire les inférences au sujet des tailles de population. Nous proposons deux types d'estimateurs de variance et d'intervalle de confiance partiellement fondés sur le plan de sondage. L'un d'eux est obtenu en utilisant un bootstrap et l'autre, en suivant la méthode delta sous l'hypothèse de normalité asymptotique. Les résultats d'une étude par simulation indiquent que (i) quand les probabilités de nomination ne sont pas faibles, chacun des ensembles d'estimateurs proposés donne de bon résultats et se comporte de façon fort semblable aux estimateurs du maximum de vraisemblance, (ii) quand les probabilités de nomination sont faibles, l'ensemble d'estimateurs dérivés en utilisant des lois a priori de Poisson donne encore des résultats acceptables et ne présente pas les problèmes de biais qui caractérisent les estimateurs du maximum de vraisemblance et (iii) les résultats précédents ne dépendent pas de la taille de la population couverte par la base de sondage.

Mots clés : Approche bayésienne; capture-recapture; approche fondée sur le plan de sondage; population finale; population d'accès difficile; maximum de vraisemblance; approche fondée sur un modèle; base de sondage.

1. Introduction

L'échantillonnage par dépistage de liens (EDP) s'est avéré être une méthode convenant bien à l'échantillonnage de populations humaines cachées ou d'accès difficile, comme les drogues, les sans abri ou les travailleurs clandestins. Il consiste à sélectionner un échantillon initial de personnes parmi la population cible et de demander à ces personnes de nommer d'autres membres de la population. Les personnes nommées qui ne figurent pas déjà dans l'échantillon initial sont alors incluses dans l'échantillon et il peut leur être demandé de nommer d'autres personnes. Ce processus se poursuit jusqu'à ce qu'une règle d'arrêt préalable soit satisfait (pour une revue de l'échantillonnage par dépistage de liens, voir Spreen 1992, ainsi que Thompson et Frank 2000). Bien que l'échantillonnage par dépistage de liens permette à l'échantillonneur de faire des inférences fondées sur un modèle valides au sujet d'un certain nombre de paramètres de population, en pratique, les hypothèses concernant l'échantillon initial sont difficiles à satisfaire. (Voir Snijders

1992, Frank et Snijders 1994, et Heckathorn 2002). Par exemple, Frank et Snijders (1994) ont élaboré une variante de l'échantillonnage par dépistage de liens dans laquelle l'échantillon initial est un échantillon de Bernoulli, c'est-à-dire que les éléments de l'échantillon initial sont sélectionnés indépendamment et avec probabilités égales; toutefois, dans les études réelles, le recrutement initial est généralement réalisé au moyen de dossiers de personnes fournis par des centres de soins de santé et des postes de police, ce qui introduit un biais de sélection appelé biais institutionnel. La difficulté à satisfaire les hypothèses au sujet de l'échantillon initial dans les situations pratiques ont poussé Félix-Medina et Thompson (2004) à élaborer une variante de l'échantillonnage par dépistage de liens qui ne nécessite pas d'échantillon de Bernoulli initial. Ils supposent qu'une population cible est couverte par une base de sondage constituée d'emplacements accessibles où les membres de la population peuvent être trouvés avec une forte probabilité (par exemple, bars, hôpitaux, îlots d'habitations ou parcs).

Remerciements

Les auteurs remercient chaleureusement les deux arbitres et l'éditeur associé qui ont grandement contribué à améliorer la lisibilité de ce texte.

Bibliographie

Deville, J.-C. (1999). Les enquêtes par panel : En quoi diffèrent-elles des autres enquêtes ? suivi de : Comment attirer une population en se servant d'une autre. *Actes des journées de méthodologie statistiques, INSEE Méthodes*, 84-85-86, 63-82.

Deville, J.-C., et Lavallée, P. (2006). Sondage indirect : Les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, 32, 2, 185-196.

Deville, J.-C., Lavallée, P. et Maunay, M. (2005). Composition, factorisation et conditions d'optimalité (faible, forte) dans la méthode de partage des poids. Application à l'enquête sur le tourisme en Bretagne. *Actes des journées de méthodologie statistiques, INSEE Méthodes*.

Hartley, H.O. (1962). Multiple Frame Surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhya*, Series C, 36, 99-118.

Huygens, C. (1673). *Horologium Oscillatorium sive de motu pendulorum*.

Deville et Maunay-Bertrand : Extensions de la méthode d'échantillonnage indirect et son application au tourisme

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.

Lavallée, P. (2002). Le sondage indirect, ou la méthode généralisée du partage des poids. Editions de l'Université de Bruxelles, éditions Ellipses, Bruxelles.

Lavallée, P., et Caron, P. (2001). Estimation par la méthode généralisée du partage des poids : Le cas du couplage d'enregistrements. *Techniques d'enquête*, 27, 171-187.

Lund, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 282-288.

Torres Manzancera, E., Sustacha Melijos, I., Menéndez Estébanez, J.M. et Valdés Peláez, L. (2002). A solution to problems and disadvantages in statistical operations of surveys of visitors at accommodation establishments and at popular visitors places. (Ed. Akos Probalid). *Proceedings Of The Sixth International Forum On Tourism Statistics. Hungarian Central Statistical Office, Budapest*.

Valdés, L., De La Ballina, J., Aza, R., Loredó, E., Torres, E., Estébanez, J.M., Dominguez, J.S. et Del Valle, E. (2001). A methodology to measure tourism expenditure and total tourism production at the regional level. Dans *Tourism Statistics : International Perspectives and Current Issues*, (Ed. Lennon, J.J.), Continuum, Grande Bretagne, 317-334.

$$\text{Var}\left[\hat{\bar{Y}}\right] = \bar{Y}^2 \text{Var}\left[\hat{F}^p\right] + T^p \text{Var}\left[\bar{Y}\right]$$

$$+ \text{Var}\left[\hat{T}^p \bar{Y}\right] \left[\text{Var}\left[\bar{Y}\right]\right]$$

+ termes liées à la non

indépendance éventuelle

des variables \hat{T}^p et \bar{Y} .

9. Illustration numérique

Un compteur mécanique d'un site en rase campagne donne $T_p = 100$ voitures. On suppose qu'il y a 20 % de voitures à une personne, 20 % de voitures à deux personnes, 20 % de voitures à trois personnes, 20 % de voitures à quatre personnes, 20 % de voitures à cinq personnes. Ainsi, la variance S_y^2 est égale à deux en négligeant les corrections de population finie. Le nombre moyen de passagers \bar{Y} est de trois. En effet, on a :

$$\bar{Y} = \frac{1}{1} \times \frac{1}{20} + \frac{2}{40} + \frac{3}{300} + \frac{1}{3} \times \frac{60}{300} = \frac{1}{80} + \frac{1}{5} \times \frac{100}{300} = \frac{4}{3}.$$

D'où $\bar{V} = 3$.
Calculons maintenant une estimation de $S_{\bar{Y}}^{1/n}$. Après simplifications de (8.8) et en supposant que T_p est suffisamment grand devant un, on a

$$S_{\bar{Y}}^{1/n} \approx \frac{1}{1} \sum_{k \in U_p} \frac{n_k}{1} - \left(\frac{\bar{Y}}{1} \right)^2.$$

Ainsi, on a

$$S_{\bar{Y}}^{1/n} = \frac{1}{30} \left(2 + 1 + \frac{3}{2} + \frac{2}{1} + \frac{5}{2} \right) - \frac{1}{3^2} = \frac{1}{30} \left(\frac{60 + 30 + 20 + 15 + 12}{30} \right) - \frac{1}{3^2} = \frac{137}{30^2} - \frac{1}{3^2} = \frac{1}{30^2}.$$

Puisque nous connaissons $S_{\bar{Y}}^{1/n}$, nous pouvons calculer la variance de l'estimateur \bar{Y} . Ainsi on a

$$\text{Var}\left[\hat{\bar{Y}}\right] \approx 3^4 \times \frac{1}{37} \times \frac{m}{1}.$$

Enfin, on peut calculer la variance de l'estimateur \hat{T}^p

Donc, afin que l'estimateur \hat{T}^p ait la même variance que l'estimateur \hat{T}^p , il suffit que la taille m de l'échantillon s_p soit égale à

$$m = 1,66n.$$

En première conclusion, on peut dire que la seconde approche rend les opérations de terrain plus simples et moins coûteuses en termes de personnel car elle ne nécessite sans contact pour obtenir la composition du ménage touristique. Elle ne nécessite qu'un échantillon environ une fois et demie plus gros que la première approche pour apporter la même précision, ce qui est tout à fait tolérable vu la simplification de la collecte qui en résulte. En pratique donc, sur tous les sites on appliquera de préférence la seconde méthode.

Conclusion

Cet article a présenté les grandes lignes d'une nouvelle méthode applicable à la statistique du tourisme. Elle consiste à saisir les touristes à partir de la consommation de certains services sur lesquels on sait construire des échantillons probabilistes. La méthode de partage des poids permet de passer de l'exactitude statistique sur les services à l'exactitude sur les unités statistiques pertinentes en tourisme : le voyage, le séjour, le ménage touristique, le touriste ou la nuitée. Cependant la méthode requiert de nombreuses adaptations et compléments au partage des poids. On s'est attardé à l'une d'elles qui est l'estimation du nombre de visiteurs d'un site en rase campagne. Deux méthodes pouvaient être mises en concurrence. L'une, plus précise en terme de taille d'échantillon, demande en fait une organisation relativement lourde et fait courir le risque d'erreurs de mesures désagréables. Au prix d'une collecte de données un peu plus abondante, on préfère donc la seconde méthode. D'autres études de ce genre ont été faites avant et pendant la réalisation de l'enquête, de sorte que la méthodologie complète est difficile à résumer en un seul article.

$$T_p = \sum_{i \in U_p} 1.$$

Le nombre moyen de passagers dans une voiture V peut s'exprimer sous la forme

$$(8.1) \qquad \overline{V} = \frac{\sum_{i \in U_p} v_i}{\sum_{k=1, \dots, K} \kappa t_k} = \frac{\sum_{i=1, \dots, N} \sum_{k=1, \dots, K} t_k}{\sum_{k=1, \dots, K} M_k / \kappa},$$

où t_k est le nombre de voitures à κ passagers et M_k le nombre de personnes venues dans une voiture à κ passagers.

Cette dernière relation permet de donner une dernière écriture de T_p

$$(8.2) \qquad T_p = T_V \overline{V}.$$

Par conséquent un estimateur de T_p s'écrit sous la forme suivante

$$(8.3) \qquad \hat{T}_p = T_V \hat{V},$$

où le nombre total de voitures T_V est parfaitement connu. En observant cette expression, on constate que pour connaître l'estimateur \hat{T}_p , il suffit de déterminer la quantité \hat{V} .

Introduisons alors l'estimateur suivant de \hat{V}

$$\hat{V} = \frac{\sum_{k \in S_p} m_k}{\sum_{k \in S_p} m_k / \kappa},$$

où m_k est le nombre de personnes de l'échantillon voyagant dans une voiture à κ passagers. L'estimateur \hat{V} peut s'écire également de la façon suivante

$$\hat{V} = \frac{\sum_{k \in S_p} 1}{\sum_{k \in S_p} 1/n_k}.$$

ou encore

$$(8.4) \qquad \hat{V} = \frac{m}{\sum_{k \in S_p} 1/n_k}.$$

Cette dernière égalité nous permet d'écire l'égalité suivante

$$(8.5) \qquad \frac{\hat{V}}{1} = \frac{m}{1} \sum_{k \in S_p} \frac{1}{n_k}.$$

Cette dernière quantité représente la moyenne empirique des $1/n_k$ et \hat{V} est la moyenne harmonique des n_k . On peut d'ailleurs calculer sa variance qui est égale à

$$(8.6) \qquad \text{Var} \left[\frac{\hat{V}}{1} \right] = \left(\frac{1}{m} - \frac{1}{1} \right) \frac{m}{1} \frac{T_p}{S_{1/n}^2}.$$

8.2 Calcul de la variance de l'estimateur de T_p sans échantillonnage de voitures

Reste à calculer la variance de l'estimateur \hat{V} sachant (8.6). Pour cela, remarquons que l'on peut écrire

$$\frac{\hat{V}}{1} = \frac{\overline{V}}{1} = \frac{\frac{1}{N} \times \left(\frac{1}{1} + \frac{\overline{V}}{1} \right)}{\frac{1}{1} + \frac{\overline{V}}{1}}$$

$$= \frac{1}{1} \left(1 - \frac{\overline{V}}{1} + \frac{\overline{V}}{1} \right) = \frac{1}{1} \left(\frac{\overline{V}}{1} - \overline{V} \right) + o \left(\frac{\overline{V}}{1} - \overline{V} \right).$$

Par conséquent, on obtient

$$\text{Var} \left[\frac{\hat{V}}{1} \right] \approx \left(\frac{\overline{V}}{1} \right)^2 \times \text{Var} \left[\frac{\overline{V}}{1} \right].$$

Finalement, on a

$$\text{Var} [V^{\hat{V}}] = \overline{V}^4 \times \text{Var} \left[\frac{\hat{V}}{1} \right],$$

ou encore, avec (8.6)

$$(8.7) \qquad \text{Var} [V^{\hat{V}}] \approx \overline{V}^4 \times \left(\frac{1}{1} \frac{T_p}{S^2_{1/n}} \right) \left(\frac{m}{1} - \frac{1}{1} \right) S^2_{1/n}.$$

Or par définition, la variance $S^2_{1/n}$ est égale à

$$(8.8) \qquad S^2_{1/n} = \frac{T_p}{1} \sum_{k \in U_p} \left(n_k \frac{1}{1} - \frac{\overline{V}}{1} \right)^2.$$

Comme la quantité T_p est inconnue, cette relation peut être estimée par

$$(8.9) \qquad \sum_{k \in S_p} \left(n_k \frac{1}{1} - \frac{\overline{V}}{1} \right)^2.$$

Grâce à (8.7) et à (8.9) on peut donc connaître facilement la variance de l'estimateur \hat{V} et par conséquent celle de l'estimateur \hat{T}_p et finalement celle de la variable d'intérêt \hat{Y} .

Remarques 8.1. L'estimateur \hat{T}_p est biaisé et asymptotiquement sans biais.

Remarque 8.2. Si les variables \hat{T}_p et \hat{Y} ne sont pas indépendantes alors on aurait

Après calculs, on obtient une équation du troisième degré en n qui s'écrit

$$\frac{\partial L}{\partial m}(n, m, \lambda) = (T^p - T^p S^p_y) S^p_y \left(-\frac{m}{1} \right) + T^p S^p_y S^p_y \left(-\frac{nm}{1} \right) + \lambda C^p = 0, \\ \frac{\partial L}{\partial n}(n, m, \lambda) = C^p n + C^p m - C = 0.$$

En annulant les dérivées partielles par rapport aux variables n , m , λ , on obtient

$$(7.13) \quad \lambda(C^p n + C^p m - C) = 0.$$

$$+ \lambda C^p = 0,$$

$$+ \frac{T^p}{T^p} S^p_y S^p_y - \bar{Y}^2 T^p S^p_y$$

$$+ (T^p - T^p S^p_y) S^p_y \frac{1}{1}$$

$$L(n, m, \lambda) = \left(\bar{Y}^2 - \frac{T^p}{S^p_y} \right) T^p S^p_y \frac{1}{1}$$

Cette équation du troisième degré en n admet une solution réelle que l'on peut déterminer avec des méthodes numériques.

En faisant le même raisonnement, on obtient une équation du troisième degré en m

$$\lambda C^p m^3 - \lambda C^p C m^2 - C^p S^p_y (T^p - T^p S^p_y) m + S^p_y (C(T^p + T^p S^p_y) + C^p T^p S^p_y) = 0.$$

7.2.2 Cas simplifié

Pour simplifier le calcul de la variance de l'estimateur \hat{Y} , nous pouvons faire une approximation dans l'égalité (7.10). En effet, nous pouvons supposer que le terme $1/nm$ est négligeable devant les termes $1/n$ et $1/m$. On obtient alors la transformation suivante de l'égalité (7.10)

$$Y = \left(\bar{Y}^2 - \frac{T^p}{S^p_y} \right) T^p S^p_y \frac{1}{1} + (T^p - T^p S^p_y) S^p_y \frac{1}{1} + \frac{T^p}{T^p} S^p_y S^p_y - \bar{Y}^2 T^p S^p_y - T^p S^p_y.$$

On cherche maintenant l'allocation des tailles des échantillons s_p et s_y qui minimise la variance de l'estimateur \hat{Y} pour des tailles de population T_p et T_y fixées.

On doit donc minimiser l'égalité (7.12) en n , m sous la contrainte

$$C^p n + C^p m = C.$$

On peut écrire l'équation lagrangienne

Rappelons également

$$T^p = \sum_{i \in U_p} v_i$$

Rappelons l'égalité suivante

8.1 Définition de \hat{T}^p

L'enquête (compte-t-on les bébés ?).

La méthode précédente peut s'avérer compliquée et coûteuse à réaliser sur certains sites. On peut obtenir une lecture plus simple en demandant à la personne k le nombre u_k de passagers de la voiture i qui l'a transportée. Ce nombre u_k est ici égal à v_i pour la voiture i qui a transporté la personne k . Cette méthode a en outre l'avantage d'obtenir avec précision le nombre de passagers au sens de

8. Construction d'un estimateur du nombre de visiteurs à partir d'un échantillonnage de visiteurs

$$m_{\text{opt}} = \frac{\left(C^p + \sqrt{C^p \frac{T^p S^p_y (T^p - T^p S^p_y)}{T^p S^p_y (T^p - T^p S^p_y)}} \right)}{C}$$

$$n_{\text{opt}} = \frac{\left(C^p + \sqrt{C^p \frac{T^p S^p_y (T^p - T^p S^p_y)}{T^p S^p_y (T^p - T^p S^p_y)}} \right)}{C}$$

Après calculs, on obtient

$$\frac{\partial L}{\partial n}(n, m, \lambda) = C^p n + C^p m - C = 0.$$

$$+ \lambda C^p = 0,$$

$$\frac{\partial L}{\partial m}(n, m, \lambda) = (T^p - T^p S^p_y) S^p_y \left(-\frac{m}{1} \right) + \lambda C^p = 0,$$

$$\frac{\partial L}{\partial n}(n, m, \lambda) = \left(\bar{Y}^2 - \frac{T^p}{S^p_y} \right) T^p S^p_y \left(-\frac{n}{1} \right) + \lambda C^p = 0,$$

Il est clair que \hat{T}_p est un estimateur sans biais du nombre total de personnes T_p et que \hat{v} estime sans biais le nombre moyen V de personnes dans une voiture.

La variance de \hat{T}_p est donc égale à

$$\text{Var}[\hat{T}_p] = T_p^2 \left(\frac{1}{n} - \frac{1}{T_p} \right) S_p^2 = \frac{1}{n} T_p^2 S_p^2 - T_p S_p^2 \quad (7.5)$$

où S_p^2 désigne la variance corrigée de la population U_p .

7.2 Construction d'un estimateur d'une variable d'intérêt dans le cas d'un échantillonnage de voitures

On veut estimer une variable d'intérêt Y de la population U_p qui s'écrit sous la forme

$$Y = \sum_{k \in U_p} Y_k, \quad (7.6)$$

où Y_k est la variable d'intérêt qu'on mesure dans le questionnaire final. Soit \hat{Y} le π -estimateur défini par

$$\hat{Y} = \sum_{k \in s_p} w_k^p Y_k, \quad (7.7)$$

où le poids w_k^p est égal à T_p / m . Par conséquent l'estimateur \hat{Y} peut s'écrire

$$\hat{Y} = \frac{\hat{T}_p}{T_p} \sum_{k \in s_p} Y_k = \hat{T}_p \bar{y} \quad (7.8)$$

en posant

$$\bar{y} = \frac{1}{m} \sum_{k \in s_p} Y_k.$$

Par la suite, les variables \hat{T}_p et \bar{y} seront supposées indépendantes. L'hypothèse est réaliste, car sur le terrain nous avons recours à deux enquêteurs indépendants.

7.2.1 Calcul de la variance de l'estimateur \hat{Y}

D'après le théorème de Huygens (1673), en conditionnant selon l'échantillon s_{Y^p} , on obtient

$$\begin{aligned} Y^p &= \text{Var}[Y] \\ &= \hat{Y}^2 \text{Var}[\hat{T}_p] + T_p^2 \text{Var}[\bar{y}] \\ &\quad + \text{Var}[\hat{T}_p] \text{Var}[\bar{y}]. \end{aligned} \quad (7.9)$$

Dans le cas présent, on assimile l'échantillon à un sondage aléatoire simple sans remise. L'égalité (7.9) devient alors

$$\begin{aligned} \frac{\partial n}{\partial \lambda}(n, m, \lambda) &= \left(\bar{Y}^2 - \frac{T_p^2}{S_p^2} \right) T_p^2 S_p^2 \frac{1}{n} \\ &\quad + T_p^2 S_p^2 S_p^2 \left(-\frac{nm}{1} \right) \\ &\quad + \lambda C_p = 0, \end{aligned}$$

En annulant les dérivées partielles par rapport aux variables n, m, λ , on obtient

$$\begin{aligned} L(n, m, \lambda) &= \left(\bar{Y}^2 - \frac{T_p^2}{S_p^2} \right) T_p^2 S_p^2 \frac{1}{n} \\ &\quad + (T_p^2 - T_p^2 S_p^2 S_p^2) \frac{1}{m} \\ &\quad + T_p^2 S_p^2 S_p^2 \frac{1}{T_p} + \frac{nm}{T_p} \\ &\quad - \bar{Y}^2 T_p^2 S_p^2 - T_p^2 S_p^2 \\ &\quad + \lambda(C_p n + C_p m - C). \end{aligned} \quad (7.11)$$

On peut écrire l'équation lagrangienne

$$C_p n + C_p m = C,$$

où C_p désigne le coût (en temps par exemple) des questionnaires posés autour des voitures, C_p le coût (en temps) des questionnaires posés aux personnes et C le coût total.

On doit donc minimiser l'égalité (7.10) en n, m sous la contrainte

$$\begin{aligned} Y^p &= \left(\bar{Y}^2 - \frac{T_p^2}{S_p^2} \right) T_p^2 S_p^2 \frac{1}{n} \\ &\quad + (T_p^2 - T_p^2 S_p^2 S_p^2) \frac{1}{m} \\ &\quad + T_p^2 S_p^2 S_p^2 \frac{1}{T_p} + \frac{nm}{T_p} \\ &\quad - \bar{Y}^2 T_p^2 S_p^2 - T_p^2 S_p^2. \end{aligned}$$

avec $S_p^2 = 1 / (T_p - 1) \sum_{k \in U_p} (Y_k - \bar{Y})^2$. En réorganisant les termes, on obtient

$$\begin{aligned} Y^p &= \bar{Y}^2 \left(\frac{1}{n} T_p^2 S_p^2 - T_p^2 S_p^2 \right) \\ &\quad + T_p^2 \left(\frac{1}{S_p^2} - \frac{T_p^2}{S_p^2} \right) \frac{m}{n} \\ &\quad + \left(\frac{1}{n} T_p^2 S_p^2 - T_p^2 S_p^2 \right) \left(\frac{1}{m} S_p^2 - \frac{T_p^2}{S_p^2} \right), \end{aligned}$$

La seconde approche utilise un échantillon de visiteurs et est destinée à estimer la même quantité à partir de l'individu interrogé qui donne le nombre de personnes qui voyagent avec lui dans la voiture. Ces deux approches sont dévoloppées dans les sections 7 et 8 suivantes.

7. Construction d'un estimateur du nombre de visiteurs à partir d'un échantillonnage de voitures

Dans ce paragraphe, on est dans le cas où un enquêteur relève en « batonnant » (c'est le terme utilisé par les praticiens du tourisme) le nombre d'occupants des voitures, c'est-à-dire, relève le nombre de personnes dans une voiture qui franchissent l'endroit où un oeil électronique ou un système équivalent a été placé pour compter les voitures dont le nombre total noté T_Y est connu avec une erreur de mesure négligeable près.

7.1 Définition et variance de $T_{A_i}^p$

Le nombre total de voitures vaut

$$(7.1) \quad T_Y = \sum_{k=1, \dots, l \in U_Y} t_k = \sum_{k \in U_Y} t_k$$

où t_k représente le nombre de voitures transportant k personnes et U_Y l'univers des voitures.

Remarques 7.1. Dans un souci d'allègement des notations, on utilisera ici et jusqu'à la fin de cet article, T_p pour $T_{A_i}^p$.

Le nombre total de personnes visitant le site vaut

$$(7.2) \quad T_p = \sum_{k=1, \dots, k \in U_p} k t_k = \sum_{k \in U_p} k t_k$$

où U_p désigne l'univers des personnes. On a aussi l'égalité

$$(7.3) \quad T_p = \sum_{l \in U_Y} v_l$$

où v_l est le nombre de personnes dans la voiture l .

Comme on l'a mentionné dans la section précédente, le nombre total de personnes T_p est inconnu. Par conséquent construisons un estimateur de T_p . Soit \hat{T}_p le π -estimateur fondé sur s_Y un échantillon aléatoire simple de voitures de taille n et de probabilité d'inclusion n/U_Y

$$(7.4) \quad \hat{T}_p = \frac{n}{T_Y} \sum_{l \in s_Y} v_l = T_Y \hat{v},$$

en posant

$$\hat{v} = \frac{1}{n} \left(\sum_{l \in s_Y} v_l \right).$$

On est ramené à une estimation sur la population des ménages touristiques. Cette formule n'est autre que celle donnée par la MGPP évoquée dans la section 2. Notons que $U^A = U^A \cup U^A \cup U^A = \bigcup_{i=1}^3 U^A$, $\theta_{AB}^H = 1$ si le service j a été utilisé par le voyage i et enfin $\delta_j = 1/\pi_j^i$.

L'estimation de la variance est possible selon les mêmes principes (cf. Lavallée (2002)). Elle ne sera pas détaillée ici car elle n'est qu'une application assez lourde en calcul des principes généraux.

De même, l'utilisation des informations auxiliaires sous forme de totaux, que ce soit dans les populations U^A ou dans la population U^B , ne pose pas de problème particulier que ce soit pour l'estimation ponctuelle ou pour l'estimation de la variance (cf. Lavallée (2002)).

Remarques 5.1. La procédure qui vient d'être décrite pour partager les poids peut-être considérée comme naïve. De fait, on sait optimiser la matrice de liens Θ_{AB} comme il est montré dans Deville et Lavallée (2006). L'application de l'enquête bretonne est décrite dans Deville, Lavallée et Maumy (2005).

6. Un exemple de problème particulier : Les points de visite en rase campagne

Comme on l'a déjà signalé, la mise en place de l'enquête sur le tourisme en Bretagne a nécessité d'assez nombreuses recherches complémentaires. On a déjà signalé ce qui concerne l'optimisation du partage des poids. L'utilisation d'informations auxiliaires relatives aux diverses bases et aux divers degrés de sondage est un autre chantier. On voudrait insister ici sur celui de l'estimation de certaines de ces informations auxiliaires, en particulier pour ce qui concerne les visites des sites touristiques en rase campagne.

Dans certains cas, on ne connaît malheureusement pas le nombre total de personnes, noté $T_{A_i}^p$, venant sur le site à un jour donné. En effet, dans l'ensemble A_i , on ne connaît pas tous les services (ici le nombre de visites) de la population. On ne peut donc pas avoir directement π_j^i et donc δ_j pour $j \in A_i$. Pour contourner ce problème, on estime alors le nombre de visiteurs journaliers afin de déduire π_j^i = $n_{A_i}/T_{A_i}^p$.

Dans la suite, on va développer deux approches d'estimation du nombre de visiteurs journaliers pour des sites accessibles en voitures uniquement (ou presque !). La première se base sur un système d'échantillonnage de voitures destiné à estimer le nombre de visiteurs sur le site.

chaque base de sondage comme une strate à la condition de pouvoir identifier, pour chaque unité échantillonnée, l'ensemble des bases dans laquelle elle figure. Elle fournit alors une solution rigoureuse, efficace et uniquement basée sur le plan à ce problème. Cette remarque pourrait fonder un article autonome, mais les auteurs savent que cela n'en vaut pas la peine : une idée qui s'exprime en dix lignes n'a pas besoin d'un article ou d'un livre pour sa survie.

4. Les paramètres d'intérêt

On définit l'application F_j qui à tout service j durant la période de référence P dans les trois types d'établissements du champ de l'enquête, associe le voyage i utilisateur de ce service.

$$F : \text{services} \rightarrow \text{voyage} \\ j \rightarrow F(j) = i.$$

Soit U^B , la population des voyages i de la période de référence P . Cette population d'intérêt U^B est l'image par F de l'ensemble des services durant la période de référence P dans les trois types d'établissements du champ de l'enquête. La population U^A est l'image par F^{-1} de l'ensemble des voyages durant la période de référence P . Pour tout $i \in U^B$, on définit $R_i(B) = \text{card}(F^{-1}(i))$, le nombre d'antécédents de i au cours de la période d'enquête, c'est-à-dire, le nombre de services j utilisés par le ménage touristique i donné.

Les paramètres d'intérêt peuvent être des totaux, des effectifs ou des ratios. Supposons par exemple, que l'on s'intéresse à l'estimation d'un total relatif à une variable y définie sur la population U^B ,

$$Y^B = \sum_{i \in U^B} y_i. \tag{4.1}$$

Un cas particulier de ces totaux est l'effectif de U^B , noté N^B et défini par

$$N^B = \text{card}(U^B) = \sum_{i \in U^B} 1.$$

Par exemple, Y^B peut-être le nombre de personnes ayant pratiqué une certaine activité, le budget total dépensé par le ménage touristique à l'intérieur de la Bretagne, la provenance géographique des ménages touristiques, le nombre de jours que le ménage touristique passe en Bretagne. Il faut noter que pour beaucoup de variables, le total Y^B dépend de la taille du ménage touristique, c'est-à-dire le nombre de personnes qui forment ce groupe et de la longueur du séjour (uniquement les jours passés en Bretagne).

Désormais, on peut écrire

ou

$$Y^B = \sum_{i \in U^B} Y_i = \sum_{j \in C_{A_j}} \sum_{i \in A_j} \sum_{i \in B_j} z_j = \sum_{i \in F^{-1}(i)} R_i(B), \text{ pour } j \in F^{-1}(i). \tag{4.2}$$

- A_1 : l'ensemble des boulangeries du champ de l'enquête repéré par l'indice a_1
- A_2 : les 16 lieux de passage du champ de l'enquête repérés par l'indice a_2
- A_3 : le péage autoroutier de La Gravelle repéré par l'indice a_3
- D_1 : l'ensemble des jours d'enquête, repérés par l'indice a_1 dans un établissement a_1 de A_1 , pour l'variant de 1 à 3
- C_{a_1} : l'ensemble des services dans un établissement a_1 de A_1 de la journée a_1 de D_1 repérés par l'indice j .

5. Estimation sans biais d'un total

Dans le paragraphe précédent, on a montré que le total d'intérêt s'écrit comme un total sur l'ensemble des services du champ. Supposons que l'on dispose d'un échantillon de δ_j sondages. Ces poids sont supposés sans biais car l'échantillon de services suit les canons d'un échantillon à plusieurs degrés, chaque sondage élémentaire étant sans biais. Pour alléger les notations, on ne fait pas apparaître, dans ce qui suit, tous les degrés de tirage de l'échantillon en fonction de l'établissement a_i . Soient

- s^B : l'ensemble des ménages touristiques i correspondant à l'ensemble des services échantillonnés au pondant à l'ensemble des services échantillonnés au cours de la période d'enquête
- s_{A_1} : l'ensemble des établissements échantillonnés
- s_{D_1} : l'ensemble des jours échantillonnés dans l'établissement a_i
- s_{a_1} : le sous-échantillon de services j correspondant au jour de l'établissement a_1 .

Disposant d'un jeu de poids de sondage δ_j pour les services répondants, et si on connaît les $R_i(B)$, on estime alors le total Y^B sans biais par

$$Y^B = \sum_{i \in s^B} w_i Y_i \tag{5.1}$$

Dans la première strate, on réalise un échantillon à trois degrés :

- un échantillon de boulangeries;
- un échantillon de jours d'enquête;
- un échantillon de clients dans la boulangerie à un jour donné.

Dans la deuxième strate, on réalise un échantillon à deux degrés :

- un échantillon de jours d'enquête;
- un échantillon de personnes qui passent sur un des 16 sites réitérés à un jour donné.

Enfin dans la troisième strate, on réalise un échantillon à deux degrés :

- un échantillon de jours d'enquête;
- un échantillon de personnes qui passent au péage autoroutier de La Gravelle à un jour donné.

On admet que tout ménage touristique consomme au moins un des « services » (achats en boulangeries, visites de sites emblématiques de la Bretagne, passage au péage autoroutier de La Gravelle), ou tout du moins, que très peu de ménages ne consomment aucun d'entre eux.

Chaque échantillonnage (boulangerie, jours, « service ») requiert des techniques particulières et il serait très long de détailler chacune d'elles. On donnera néanmoins les quelques indications techniques suivantes :

- les boulangeries sont échantillonnées selon un plan classique stratifié géographiquement (cinq strates : partie « littorale » des quatre départements bretons, intérieur de la Bretagne). Dans chaque strate les boulangeries sont échantillonnées avec des probabilités proportionnelles à leur « potentiel touristique » construit à partir de leur chiffre d'affaire et de la capacité d'hébergement touristique et du nombre de résidences principales de la commune à laquelle elles appartiennent. Théoriquement du moins, car pratiquement ce tirage a été un peu « forcé » par des circonstances fortuites (refus de boulangers, fermetures durant certaines périodes par exemple).
- Les sites ne sont pas échantillonnés mais choisis pour leur notoriété et pour la possibilité technique d'y définir un « point de passage obligé » (parfois appartenant à la commune à laquelle elles appartiennent).
- Pour chaque boulangerie, chaque site et pour le péage autoroutier de La Gravelle, on a défini des « grappes de jours » complètement homogènes de chaque période P. Une grappe a été attribuée aléatoirement à chaque boulangerie, site ainsi qu'au péage autoroutier de La Gravelle. Pratiquement cela signifie qu'un enquêteur employé à plein temps est mobilisé sur plusieurs grappes.

— Pour chacun des « services » les utilisateurs sont échantillonnés selon des techniques habituelles de sélection aléatoire à mesure des arrivées : échantillon pseudo-systématique car pendant que l'enquêteur fait accepter un questionnaire, des gens passent sans qu'il puisse compter. Le nombre total de visiteurs ne peut donc pas être estimé directement. Si un site est accessible par une billetterie (musée ou château par exemple) l'échantillonnage s'appuie sur elle. Au final, l'échantillon d'utilisateurs d'un « service » à un jour donné est considéré comme un échantillon bernoullien, c'est-à-dire un sondage aléatoire simple si on connaît la taille de la population c'est-à-dire le nombre de visiteurs du jour donné.

Remarques 3.1. La définition même du *touriste* est liée à l'hébergement, et il paraît naturel d'utiliser une base directement liée à ce service. La pratique montre que c'est difficilement réalisable.

On n'a, d'abord, aucune base de sondage correcte pour l'hébergement non marchand (parents, amis, résidence secondaire) ni pour les locations meublées saisonnières.

Pour l'hébergement en hôtels, campings et gîtes familiaux, les tests de l'été 2004 ont montré l'existence de biais catastrophiques liés à l'intervention des hôteliers dans le processus de sélection des enquêtes. Ceux-ci ne respectent absolument pas les consignes d'échantillon aléatoire et distribuent « essentiellement » les questionnaires à leurs bons clients. Cette partie du dispositif de l'enquête a dû être abandonnée et remplacée par le passage au péage autoroutier de La Gravelle, qui est régulièrement l'objet d'enquêtes de qualité honnête faites par divers organismes.

Par ailleurs, les questionnaires collectés dans les boulangeries et sur les sites emblématiques de la Bretagne pendant l'été 2004, rendent apparemment (qualitativement et quantitativement) bien compte des différents modes d'hébergement. De même, l'alimentation eut sans doute mieux été capturée par des questionnaires à la sortie des supermarchés. Mais là, le problème réside dans l'hétérogénéité de ces établissements et dans la lutte au couteau que se livrent les enseignes, le groupe C... accepte les enquêtes dans ses établissements uniquement si le groupe I... en est exclu ! En revanche, l'adhésion des artisans boulangers au concept de l'enquête a été excellente.

Remarques 3.2. Par définition même de la méthode utilisée, on se place formellement dans le contexte de l'échantillonnage à partir de bases multiples. Le problème a donné lieu à une abondante littérature (Hartley (1962), Lund (1968) et Hartley (1974) pour le moins). La MGFP s'applique à ce problème en considérant tout simplement

comme l'hébergement, la nourriture, les activités de loisirs, les transports.

3.2 La population d'intérêt

Soit G un champ géographique (les quatre départements bretons) et P une période de référence (pour nous celle qui s'étend du mois de février 2005 au mois de décembre 2005).

Un touriste est une personne ayant passé au moins une nuit dans G hors de sa résidence principale (nuitée). Pour un touriste, un séjour est un intervalle sej de P de durée le cardinal de sej noté $|sej|$, au cours duquel le touriste passe toutes ses nuits dans G hors résidence principale et, les nuits immédiatement avant ou après le séjour sej étant passées hors de G (ou à la résidence principale).

Un voyage est un ensemble de touristes (ménage touristique) partageant le même séjour et avec le même hébergement au cours du séjour. On utilisera aussi le terme de ménage touristique pour faire plusieurs voyages au cours d'une période, mais nous n'avons aucun moyen de les distinguer).

L'unité statistique i de l'enquête est le voyage. Les sous unités d'enquête sont les séjours, les touristes et les nuitées. Un voyage i comporte n_i touristes pendant le séjour de durée $|sej|$ et donc $n_i \times |sej|$ nuitées. Ici la population U^B est donc l'ensemble des voyages dans G au cours de P . ($sej \cap P \neq \emptyset$).

3.3 Le plan de sondage de l'enquête

Pour utiliser la MGPP, la population théorique U^A est constituée par un ensemble de « services ». Dans cette en-

quête, ceux-ci sont constitués par :
— les achats en boulangerie, constituant une première strate de U^A .
— les visites d'un ensemble de sites culturels ou de loisirs ou familiaux très connus. En pratique, pour chacun d'eux, un « point de passage obligé » a été défini. C'est l'ensemble des passages par ce point qui est la seconde strate de U^A .

— les passages sortant de la Bretagne au péage autoroute de La Gravelle qui regroupe environ 80 % des sorties des touristes de la Bretagne en voiture. Ce mode de transport caractérise lui-même 80 % des jours de non-résidents bretons. Ce passage constitue la troisième strate de U^A .

En d'autres termes, la base de sondage est donc formellement constituée de trois strates :

1. les achats en boulangerie;
2. les visites d'un ensemble de sites emblématiques de la Bretagne;
3. le passage au péage autoroutier de La Gravelle.

Pour chaque unité i de s^B , une variable d'intérêt y_i est mesurée.

On suppose que, pour toute unité j de s^A , on peut obtenir les valeurs de θ_{AB}^{ji} pour $i = 1, \dots, N_B$ par entrevue directe ou à partir d'une source administrative. Pour toute unité i identifiée de U^B (ou seulement de s^B), on suppose que l'on peut obtenir les valeurs de θ_{AB}^{ji} pour $j = 1, \dots, N_A$. Par conséquent, il n'est pas nécessaire de connaître les valeurs de θ_{AB}^{ji} pour la totalité de la matrice de liens Θ_{AB} . En fait, on ne doit connaître les valeurs de θ_{AB}^{ji} que pour les lignes j de Θ_{AB} , où $j \in s^A$, ainsi que pour les colonnes i de Θ_{AB} où $i \in s^B$.

Par exemple si le but est d'estimer une variable d'intérêt y^B de la population cible U^B , où

$$y^B = \sum_{i=1}^{N_B} w_i y_i, \quad (2.1)$$

où w_i est le poids d'estimation de l'unité i de s^B , avec $w_i = 0$ pour $i \notin s^B$. Pour obtenir une estimation sans biais d'une variable d'intérêt y^B , il suffirait d'utiliser comme poids w_i . L'inverse de la probabilité de sélection π_i^j de l'unité i . Comme il est mentionné dans Lavallée (1995) et Lavallée (2002), il est généralement difficile, voire impossible, d'obtenir ces probabilités. On a alors recours à la MGPP. Dans celle-ci les poids sont donnés par

$$w_i = \sum_{j \in s^A} \frac{\theta_{AB}^{ji}}{w_j^A},$$

où $\theta_{AB}^{ji} = \theta_{AB}^{ji} / \sum_{i=1}^{N_B} \theta_{AB}^{ji}$. De cette construction, l'estimateur \hat{y}^B est sans biais. De même, la variance de cet estimateur peut-être calculée et estimée car elle est identique à celle de

$$\sum_{j \in s^A} \frac{w_j^A}{z_j^A} \sum_{i=1}^{N_B} \theta_{AB}^{ji} y_i,$$

3.1 Objectifs de l'enquête

3. L'enquête tourisme en milieu ouvert

Le principe de l'enquête est le suivant :
« atteindre les touristes (étrangers ou français habitant la Bretagne ou pas) par le biais de services destinés à satisfaire leurs besoins élémentaires ou spécifiques »

Extensions de la méthode d'échantillonnage indirect et son application

Jean-Claude Deville et Myriam Maunay-Bertrand

Résumé

On doit procéder à une enquête portant sur la fréquentation touristique d'origine intra ou extra-régionale en Bretagne. Pour des raisons matérielles concrètes, les « enquêtes aux frontières » ne peuvent plus s'organiser. Le problème majeur est l'absence de base de sondage permettant d'atteindre directement les touristes. Pour contourner ce problème, on applique la méthode d'échantillonnage indirect dont la pondération est obtenue par la méthode généralisée de partage des poids développée récemment par Lavallée (1995), Lavallée (2002), Deville (1999) et présentée également dans Lavallée et Caron (2001). Cet article montre comment adapter cette méthode à l'enquête. Certaines extensions s'avèrent nécessaires. On développera l'une d'elle destinée à estimer le total d'une population dont on a tiré un échantillon bernoullien.

Mots clés : Méthode généralisée de partage des poids ; base incomplète et bases multiples.

1. Introduction

Une « enquête aux frontières » portant sur la fréquentation touristique extra-régionale en Bretagne (hormis celle des Bretons) a été réalisée sur la période d'avril à septembre 1997. L'Observatoire Régional du Tourisme de Bretagne et les Comités Départementaux de Tourisme aimeraient recommencer ce type d'enquête. Malheureusement ils n'ont plus la possibilité de recueillir une certaine masse d'informations récoltées aux frontières régionales ou intra-régionales, car les forces de police ne désirent plus collaborer à la réalisation d'enquêtes au bord des routes.

C'est pourquoi l'Observatoire Régional du Tourisme de Bretagne avec l'aide d'un comité technique constitué de méthodologues et d'opérateurs de terrain ont décidé de mettre en place une nouvelle méthodologie d'enquête en remplaçant de la méthodologie des « enquêtes aux frontières ». De plus, l'évaluation de la part du tourisme intra-régional (des Bretons prenant des vacances en Bretagne, par exemple) est indispensable pour définir les facteurs de développement.

Un des problèmes majeurs est l'absence d'une base de sondage permettant d'interroger directement les touristes. Pour contourner ce problème, l'idée principale, déjà utilisée par la région des Asturies en Espagne (Valdes, De La Ballina, Azza, Loredó, Torres, Estebanez, Domínguez et Del Valle (2001) et Torres Manzanaera, Susaccha Meljosa, Menéndez Estebanez et Valdes Pelaez (2002)), est d'échantillonner des services destinés principalement aux touristes et de les interroger sur les différents lieux de ces nombreuses prestations touristiques. Il est bien évident qu'un touriste peut utiliser une ou plusieurs fois un ou plusieurs services de la base de sondage pendant la période

2. La méthode généralisée de partage des poids

On va rappeler très brièvement le principe de la méthode généralisée de partage des poids (MGPP). Pour de plus amples informations, on renvoie à Lavallée (1995), Lavallée (2002) et Deville (1999).

Soient U^A une population finie contenant N^A unités, où chaque unité est désignée par j et U^B une population finie contenant N^B unités, où chaque unité est désignée par i . La correspondance entre U^A et U^B peut être représentée par une matrice de liens $\Theta_{AB} = [\theta_{ij}^{AB}]$, de taille $N^A \times N^B$ où chaque élément $\theta_{ij}^{AB} \geq 0$. Autrement dit, l'unité j de U^A est reliée à l'unité i de U^B à condition que $\theta_{ij}^{AB} > 0$; sinon, il n'existe aucun lien entre les deux unités.

Dans le cas du sondage indirect, on sélectionne l'échantillon s^A de n^A unités à partir de U^A selon un plan d'échantillonnage donné. Soit $\pi_j^A > 0$, la probabilité de sélection de l'unité j . Pour chaque unité j sélectionnée dans s^A , on identifie les unités i de U^B pour lesquelles $\theta_{ij}^{AB} > 0$. Soit s_B , l'ensemble des n^B unités de U^B identifiées au moyen des unités $j \in s^A$, c'est-à-dire

$$s_B = \{i \in U^B; \exists j \in s^A \text{ et } \theta_{ij}^{AB} > 0\}.$$

où \mathbf{H}_{G, H^T} est une matrice diagonale de taille $N_G^I \times N_G^I$ de valeur 1, « remboursée » de zéros. En suivant exactement le même scénario que l'exemple 1, un élément type de \mathbf{H}_{G, H^T} est donné par 1 si i et i' sont toutes deux liées à la même unité j de U_A (c'est-à-dire liées à l'unité g de U_G), et 0 sinon. Par conséquent nous pouvons voir facilement dans quel cas les conditions (6.7), (6.8a) ou (6.8b) peuvent être satisfaites. En fait, comme toutes les composantes de $\mathbf{\theta}_{GB, op, i'}$ sont égales, $\Delta_{GB, op, i'}$ est un vecteur proportionnel à la somme des lignes de Δ_{G, H^T} , c'est-à-dire la somme des lignes de

$$\left[\mathbf{H}_{G, H^T} - \frac{1_{G, i'} 1_{G, i'}^T N_A}{N_A} \right].$$

Mais (6.7) dit que ce vecteur doit avoir les mêmes composantes. Cela n'est possible que si et seulement si la matrice \mathbf{H}_{G, H^T} ne contient que des zéros, ou qu'elle est de dimension 1×1 , ce qui se produit lorsque i et i' sont chacune liées uniquement à un élément de U_A . Donc, comme pour l'échantillonnage de Poisson, une optimalté fort indépendante de \mathbf{Y} n'a généralement pas lieu dans le cas de l'échantillonnage aléatoire simple.

7. Conclusion

Dans le présent article, nous avons discuté de l'utilisation du sondage indirect conjugué à la méthode généralisée du partage des poids (MGPP) élaborée pour produire des poids. Puis, nous avons démontré les propriétés suivantes de la MGPP : absence de biais, calcul de la variance et transitivity. Ensuite nous avons présenté une section sur l'utilisation de la MGPP lorsque les liens entre les populations U_A et U_B sont exprimés par des valeurs 1 et 0, c'est-à-dire qu'il existe un lien ou qu'il n'en existe pas. La section suivante a été consacrée aux résultats obtenus avec diverses formes de matrices de liens. Enfin, nous avons abordé le problème de l'optimalté, c'est-à-dire le choix des valeurs optimales pour exprimer les liens entre U_A et U_B de façon à minimiser la variance des estimations obtenues en appliquant la MGPP. Nous avons fait la distinction entre deux formes d'optimisation, à savoir l'optimisation faible et l'optimisation forte. L'optimisation faible consiste à trouver les valeurs des liens qu'il convient d'utiliser pour minimiser, pour chaque unité, la variance des poids produits par la MGPP. La solution est toujours définie de façon unique, et est facile à calculer et à appliquer en pratique. L'optimisation faible est également une condition nécessaire de l'optimisation forte. L'optimisation forte consiste à trouver les valeurs des liens permettant de minimiser la variance du total

Ernst, L. (1989). Weighing issues for longitudinal household and family estimates. Dans *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York : John Wiley & Sons, Inc. 135-159.

Deville, J.C., (1998). Comment attraper une population en se servant d'une autre. *Insee méthodes*, n°84-85-86, *Actes des Journées de méthodologie statistique des 17-18 mars 1998*, 63-82.

Jazminski, A.H. (1970). *Stochastic Processes and Filtering Theory*. New York : Academic Press.

Kalton, G., et Brick, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 1, 37-49.

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 1, 27-35.

Lavallée, P. (2001). Estimation par la méthode généralisée du partage des poids : Le cas du couplage d'enregistrements. *Techniques d'enquête*, 27, 2, 171-187.

Sämdal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Searle, S.R. (1971). *Linear Models*. New York : John Wiley & Sons, Inc.

Bibliographie

Les auteurs remercient toutes les personnes qui ont manifesté un intérêt pour le sondage indirect et particulièrement la MGPP. Elles ont motivé la rédaction de cet article qui dépasse le cadre de ce qui avait été écrit antérieurement à ce sujet.

Remerciements

Nous recommandons d'utiliser l'optimisation faible, parce qu'elle coule de source et qu'elle est très facile à utiliser. En outre, si notre problème d'estimation peut être optimisé également au sens fort, nous aurons obtenu ce résultat par la voie de l'optimisation faible, même si nous ne l'avons pas démontré.

(c'est-à-dire liées à l'unité g de U^G provenant de la même unité j de U^A), et 0 autrement. Par conséquent, si deux unités i et i' ne sont pas liées aux mêmes unités de U^A , alors $\Delta_{G,ii'}$ est une matrice de zéros et les conditions (6.7), (6.8a) et (6.8b) sont automatiquement satisfaites. Si nous nous référons à la figure 1, les enfants $i = 2$ et $i' = 3$ de U^B ne sont pas apparentés aux mêmes parents j de U^A . Si la sélection des parents est faite par échantillonnage de Poisson ou de Bernoulli, la matrice $\Delta_{G,23}$ de dimension 2×2 ne contiendra alors que des zéros, c'est-à-dire

$$\Delta_{G,23} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Donc, les relations (6.7), (6.8a) ou (6.8b) seront satisfaites avec $\Phi_2 = 0$, ce qui exprime le fait que les poids de i et de i' ne sont pas corrélés.

Si deux unités i et i' sont reliées à la même unité j de

U^A , alors, si nous utilisons (6.7), le vecteur colonne $\Delta_{G,ii'}^{GB, opt, i}$ contient le scalaire $(\tau_{ii'}^G)^{-1} = [\sum_{N_j^A=1}^{N_j^A} \pi_{ii'}^G]^{-1}$ pour ses N_j^A premières composantes, et 0 pour les

$N_j^B - N_j^A$ composantes restantes (en supposant que $N_j^B \geq N_j^A$). Comme la quantité $\Delta_{G,ii'}^{GB, opt, i}$ doit être égale à $\Phi_{ii'}^{ii'} \mathbf{I}_{G,i}$ pour satisfaire (6.7), elle doit contenir

uniquement la valeur $\Phi_{ii'}^{ii'}$. Puisque $\Phi_{ii'}^{ii'} = \theta_{G,ii'}^{GB, opt, i}$, cela se produira uniquement si le vecteur

$\theta_{G,ii'}^{GB, opt, i} = [\mathbf{1}]$, ce qui signifie qu'il n'existe qu'un seul lien vers l'unité i de U^B . Comme nous le voyons, il ne s'agit pas d'une condition qui sera satisfaite en général et, par

conséquent, nous pouvons dire, dans le cas de l'échantillonnage de Poisson, il n'y aura généralement pas d'optimisation forte indépendante de \mathbf{Y} .

Pour conclure, nous pourrions dire que dans le cas de l'échantillonnage à l'unité g de U^G (6.7), (6.8a) ou (6.8b) seront satisfaites en pratique uniquement si les unités de U^A sont liées à une seule unité de U^B , comme dans le cas de l'échantillonnage des ménages en utilisant une liste de personnes. Dans les autres cas, la matrice optimale $\theta_{AB, opt}^{AB, opt}$ obtenue par optimisation faible ne donnera vraisemblablement pas lieu à une optimisation forte indépendante de \mathbf{Y} .

Exemple 2 : Échantillonnage aléatoire simple

Supposons que l'on sélectionne l'échantillon s^A par échantillonnage aléatoire simple. Dans ce cas, la matrice Δ^A de taille $N^A \times N^A$ est donnée par

$$\Delta^A = \frac{N^A}{N^A(N^A - n^A)} \left[\mathbf{I}^A - \frac{N^A}{N^A} \mathbf{1}^A \mathbf{1}^{A'} \right].$$

Si nous considérons la factorisation de la section 6.1, nous obtenons

$$\Delta_{G,ii'}^{AG} = \frac{N^A}{N^A(N^A - n^A)} \left[\mathbf{H}_{G,ii'} \times \left[\mathbf{I}_{G,i} \mathbf{1}_{G,i'}^A \right] - \frac{N^A}{N^A} \mathbf{1}_{G,i}^A \mathbf{1}_{G,i'}^A \right]$$

par

chaque sous-matrice $\Delta_{G,ii'}^{AG}$ de taille $N_G^i \times N_G^{i'}$ est donnée par $\theta_{AB, opt}^{AB, opt}$ que nous obtenons par optimisation faible. D'abord, indépendante de \mathbf{Y} , sont satisfaites pour la matrice optimale conditions (6.7), (6.8a) ou (6.8b) pour une optimisation forte

En utilisant le résultat 7, nous vérifions maintenant si les conditions (6.7), (6.8a) ou (6.8b) pour une optimisation forte sont satisfaites pour la matrice optimale $\theta_{AB, opt}^{AB, opt}$ que nous obtenons par optimisation faible. D'abord, indépendante de \mathbf{Y} , sont satisfaites pour la matrice optimale conditions (6.7), (6.8a) ou (6.8b) pour une optimisation forte

De nouveau, ce résultat est important, car il va directement dans le sens des résultats de Kalton et Brick (1995), de Lavallée (2002), et de Lavallée et Caron (2001). Autrement dit, dans le cas de l'échantillonnage aléatoire simple, le choix optimal de $\theta_{AB, opt}^{AB, opt}$ devrait être 1 s'il existe un lien entre l'unité j de U^A et l'unité i de U^B , et 0 sinon.

En utilisant (2.6), nous obtenons les poids optimaux \mathbf{W}^{opt} en utilisant (2.6).

Enfin, après avoir calculé la matrice optimale $\theta_{AB, opt}^{AB, opt}$, nous utilisons ces valeurs pour construire les vecteurs $\theta_{G,ii'}^{GB, opt, i}$ puis la matrice $\theta_{G,ii'}^{GB, opt, i}$.

Nous utilisons ces valeurs pour construire les vecteurs $\theta_{G,ii'}^{GB, opt, i}$ puis la matrice $\theta_{G,ii'}^{GB, opt, i}$.

Ensuite, partant de (6.5), nous obtenons directement les valeurs optimales

$$\Delta_{G,ii'}^{AG} = \frac{N^A}{N^A(N^A - n^A)} \left[\mathbf{I}_{G,i} \mathbf{1}_{G,i'}^A - \frac{N^A}{N^A} \mathbf{1}_{G,i}^A \mathbf{1}_{G,i'}^A \right]$$

nous obtenons

qui est de taille N_G^i . Alors, en utilisant un résultat matriciel que l'on peut trouver, entre autres, dans Jazwinski (1970), nous obtenons

$$\Delta_{G,ii'}^{AG} = \frac{N^A}{N^A(N^A - n^A)} \left[\mathbf{I}_{G,i} \mathbf{1}_{G,i'}^A - \frac{N^A}{N^A} \mathbf{1}_{G,i}^A \mathbf{1}_{G,i'}^A \right]$$

Chaque sous-matrice $\Delta_{G,ii'}^{AG}$ est donnée par

l'unité j de U^A . De $\Delta_{G,ii'}^{AG}$, nous extrayons les sous-matrices $\mathbf{I}^{A, ii'}$ est une matrice carrée de taille N_j^A , avec N_j^A égal au nombre de liens (ou de « flèches ») ayant pour origine

l'unité j de U^A . De $\Delta_{G,ii'}^{AG}$, nous extrayons les sous-matrices $\mathbf{I}^{A, ii'}$ est une matrice carrée de taille N_j^A , avec N_j^A égal au nombre de liens (ou de « flèches ») ayant pour origine

l'unité j de U^A . De $\Delta_{G,ii'}^{AG}$, nous extrayons les sous-matrices $\mathbf{I}^{A, ii'}$ est une matrice carrée de taille N_j^A , avec N_j^A égal au nombre de liens (ou de « flèches ») ayant pour origine

montrer que la variance optimale (quand elle existe) a pour

expression $\mathbf{V}'\mathbf{V}\mathbf{Y}$.

Démontrons que cet ensemble de conditions est

également suffisant. Supposons que (6.7) est vérifiée.

Notons que, pour $t = i$, la condition (6.7) n'est rien d'autre

que (6.5) qui donne les valeurs nécessaires pour les

$\theta_{GB, \text{opt}, i}^t$. Il est maintenant simple de vérifier que (6.4) tient

quelle que soit la valeur de \mathbf{Y} et que nous avons obtenu

l'optimalité forte. Les valeurs de λ_i dépendent de \mathbf{Y} , ainsi

que de la variance $\text{Var}(\mathbf{Y}_B)$, mais nous savons que les

équations (6.4) ont toujours la même solution (6.5) qui ne

dépend pas de \mathbf{Y} . Par conséquent, nous avons le résultat

suitant :

Résultat 7 :

Les conditions $\Delta_{G, H}^{GB, \text{opt}, i} = \Phi^{H'} \mathbf{1}_{G, i}$ sont nécessaires

et suffisantes pour qu'il existe une matrice de liens

normalisée $\Theta_{GB, \text{opt}, i}^t$ ou de façon équivalente, $\Theta_{AB, \text{opt}, i}^t$ qui

permet d'obtenir une optimalité forte indépendante du

vecteur \mathbf{Y} de la variable d'intérêt. Les valeurs figurant dans

les colonnes de cette matrice optimale forte sont données

par (6.5) qui sont les vecteurs $\theta_{GB, \text{opt}, i}^{GB, \text{opt}, i}$ obtenus à partir de

l'optimalité faible.

Il convient de souligner que, puisque $\Delta_{G, H}^{GB, \text{opt}, i} =$

$\lambda_i \mathbf{1}_{G, i}$, l'expression (6.7) peut s'écrire de façon équivalente

$$\Phi^{H'} \theta_{GB, \text{opt}, i}^{GB, \text{opt}, i} = \Delta_{G, H}^{GB, \text{opt}, i} \theta_{GB, \text{opt}, i}^{GB, \text{opt}, i} \quad (6.8a)$$

ou

$$\Phi^{H'} \mathbf{1}_{G, i} = \Delta_{G, H}^{GB, \text{opt}, i} \Delta_{G, H}^{-1} \mathbf{1}_{G, i} \quad (6.8b)$$

où $\Phi^{H'} = (\theta_{GB, \text{opt}, i}^{GB, \text{opt}, i} \Delta_{G, H}^{-1} \Delta_{G, H}^{GB, \text{opt}, i}) (\mathbf{1}_{G, i} \Delta_{G, H}^{-1} \mathbf{1}_{G, i})$ et $\Phi^{H'} =$

$(\theta_{GB, \text{opt}, i}^{GB, \text{opt}, i} \Delta_{G, H}^{-1} \Delta_{G, H}^{GB, \text{opt}, i}) (\mathbf{1}_{G, i} \Delta_{G, H}^{-1} \mathbf{1}_{G, i})$. Dans certains situ-

ations, ces expressions peuvent s'écrire plus faciles à

utiliser que l'expression (6.7) énoncée dans le résultat 7.

6.5 Deux exemples

Nous présentons maintenant deux exemples qui illustrent

la théorie que nous venons d'exposer sur l'optimalité faible

et l'optimalité forte indépendante de \mathbf{Y} .

Exemple 1 : Échantillonnage de Poisson

Supposons que l'échantillon s^A soit sélectionné par

échantillonnage de Bernoulli ou de Poisson. Dans ce cas, la

matrice Δ_A de taille $N^A \times N^A$ est donnée par

$\Delta_A = \text{diag}(1/\pi_A^i - 1)$. Si nous considérons la factorisation

de la section 6.1, nous avons $\Delta_{AG} = \Theta_{AG}^A \Delta_A^{AG} \Theta_{AG}^A =$

$\Theta_{AG}^{AG} [\text{diag}(1/\pi_A^i - 1)] \Theta_{AG}^A$ avec N_{AG}^A égal au

nombre de liens (ou « flèches ») ayant pour origine l'unité

f de U^A . De Δ_G , nous extrayons les sous-matrices $\Delta_{G, H}^{AG}$ qui sont, ici, diagonales. Chaque sous-matrice $\Delta_{G, H}^{AG}$ est donnée par $\Delta_{G, H}^{AG} = \text{diag}(1/\pi_A^i - 1)$, qui est de taille N_G^i . Notons que chaque valeur $(1/\pi_A^i - 1)$ correspond simplement à l'unité g de U_G^A qui a été liée à l'unité i de U^A qui a été liée antérieurement à l'unité g de U_G^A , qui a son tour a été liée à l'unité i de U^B . Maintenant, partant de (6.5), nous obtenons directement les valeurs optimales $\theta_{GB, \text{opt}, i}^{GB, \text{opt}, i}$ qui minimisent $\text{Var}(\mathbf{Y}_B)$, au sens faible. Ces valeurs sont données par les vecteurs

$$\theta_{GB, \text{opt}, i}^{GB, \text{opt}, i} = \left\{ \frac{\pi_A^i}{\pi_A^g} (1 - \pi_A^g) \tau_{G, i}^i, \dots, \frac{\pi_A^i}{\pi_A^g} (1 - \pi_A^g) \tau_{G, i}^i \right\}$$

où

$$\tau_{G, i}^i = \sum_{N_G^i} \pi_A^g / (1 - \pi_A^g), i = 1, \dots, N_B.$$

Les $\theta_{GB, \text{opt}, i}^{GB, \text{opt}, i}$ sont utilisés pour construire les vecteurs $\theta_{GB, \text{opt}, i}^{GB, \text{opt}, i}$ puis la matrice $\Theta_{GB, \text{opt}, i}^{GB, \text{opt}, i} = \{\theta_{GB, \text{opt}, i}^{GB, \text{opt}, i}, \dots, \theta_{GB, \text{opt}, i}^{GB, \text{opt}, i}\}$. Enfin, après avoir calculé la matrice optimale, $\Theta_{AB, \text{opt}, i}^{AB, \text{opt}, i} = \Theta_{GB, \text{opt}, i}^{AB, \text{opt}, i} \Theta_{GB, \text{opt}, i}^{GB, \text{opt}, i}$, nous obtenons les poids optimaux \mathbf{W}^{opt} en utilisant (2.6).

Il convient de souligner que, si les probabilités d'inclusion π_A^i sont égales, nous obtenons

$$\theta_{GB, \text{opt}, i}^{GB, \text{opt}, i} = \left\{ \frac{1}{N_G^i}, \dots, \frac{1}{N_G^i} \right\} = \frac{1}{N_G^i} \mathbf{1}_{GB, i}^{N_G^i}$$

où N_G^i est tout simplement le nombre d'unités de U^A liées à l'unité i de U^B . Autrement dit, dans le contexte de l'échantillonnage de Bernoulli (c'est-à-dire l'échantillonnage de Poisson avec probabilités égales), pour minimiser la variance $\text{Var}(\mathbf{Y}_B)$, le choix des valeurs de $\theta_{AB, i}^{AB, \text{opt}, i}$ devrait être 1 s'il existe un lien entre l'unité j de U^A et l'unité i de U^B , et 0 autrement. Cela correspond aux résultats obtenus par Kalton et Brick (1995), Lavalée (2002), ainsi que Lavalée et Caron (2001).

En utilisant le résultat 7, nous vérifions maintenant si les conditions (6.7), (6.8a) ou (6.8b) que nous avons obtenue par matrice optimale $\Theta_{AB, \text{opt}, i}^{AB, \text{opt}, i}$ que nous avons obtenue par optimisation faible. Le cas échéant, cette matrice donne aussi une optimalité forte indépendante de la variable d'intérêt \mathbf{Y} . Premièrement, nous avons

$$\Delta_{G, H}^{AG} = \text{diag} \left(\frac{\pi_A^g}{\pi_A^g} \mathbf{1} - \frac{\pi_A^g}{\pi_A^g} \right).$$

En outre, chaque sous-matrice $\Delta_{G, H}^{AG}$ de taille $N_G^i \times N_G^i$ a pour taille N_G^i de la diagonale $\Delta_{G, H}^{AG}$ est de zéro. Autrement dit, un élément typique de $\Delta_{G, H}^{AG}$ est donné par $(1/\pi_A^i - 1)$ sur une partie de la diagonale si i et j sont toutes deux liées à la même unité j de U^A .

sous-espace nul de Θ^{GB} n'est pas inclus dans l'espace nul de $\Delta^{G,i}$.

Le problème d'optimisation susmentionné peut être réécrit sous une forme différente. Soit $\Delta^{G,i}$ la sous-matrice de Δ^G correspondant aux éléments qui occupent les positions g et g' si g possède un lien avec l'unité i et que g' possède un lien avec l'unité i' . Ces matrices $\Delta^{G,i}$ sont symétriques, définies positives et que $\Delta^{G,i} = \Delta^{G,i'}$. Sous ces notations, le problème d'optimisation peut s'écrire sous la forme :

Minimiser

$$(6.3) \quad \sum_{t=1}^{I=1} \sum_{N^B} y_i^t y_{i'}^t \Theta^{GB,i} \Delta^{G,i} \Theta^{GB,i'}$$

sous les contraintes $\Theta^{GB,i} \mathbf{1}_{G,i} = 1$ pour tout $i = 1, \dots, N^B$.

La minimisation est réalisée pour les vecteurs $\Theta^{GB, \text{opt}, i}$ qui satisfont

$$(6.4) \quad y_i^t \Delta^{G,i} \Theta^{GB, \text{opt}, i} y_{i'}^t = \lambda_i^t \mathbf{1}_{G,i}$$

pour tout $i = 1, \dots, N^B$ et où les λ_i^t représentent les multiplicateurs de Lagrange entrant dans la minimisation sous contraintes de (6.3). Comme le montre (6.4), le choix optimal $\Theta^{GB, \text{opt}, i}$ (et par conséquent $\Theta^{GB, \text{opt}}$) dépend en général explicitement du vecteur \mathbf{Y} , ce qui n'est pas utile en pratique. Observons que l'ensemble des λ_i^t dépend aussi de la variable \mathbf{Y} . Ce qui apparaît plus explicitement à la section 6.3. Cette raison est celle pour laquelle, au lieu d'une optimisation forte, nous rechercherons une forme plus faible dominant une solution « optimale » ($\Theta^{GB, \text{opt}}$ (et $\Theta^{AB, \text{opt}}$) indépendante de \mathbf{Y} .

6.3 Optimabilité faible

Les équations (6.4) doivent être valides pour tout vecteur \mathbf{Y} . En particulier, une condition nécessaire est qu'elles doivent être vérifiées pour une variable d'intérêt particulière, telle que $y_i = 1$ pour une unité i de U^B et $y_{i'} = 0$ pour toutes les autres unités i' de U^B ($i' \neq i$). Cela nous donne les conditions nécessaires (une pour chacune de ces variables particulières) $\Delta^{G,i} \Theta^{GB, \text{opt}, i} = \lambda_i^t \mathbf{1}_{G,i}$. Si nous supposons que $\Delta^{G,i}$ est inversible, nous obtenons alors $\Theta^{GB, \text{opt}, i} = \lambda_i^t \Delta^{G,i-1} \mathbf{1}_{G,i}$. Il peut être démontré qu'il s'agit aussi d'une condition suffisante. Maintenant, comme $\Theta^{GB, \text{opt}, i} \mathbf{1}_{G,i} = 1$, nous avons $\lambda_i^t = 1 / \mathbf{1}_{G,i}' \Delta^{G,i-1} \mathbf{1}_{G,i}$. Par conséquent, une condition nécessaire et suffisante pour que l'équation (6.4) soit satisfaite est que

$$(6.5) \quad \Theta^{GB, \text{opt}, i} = \frac{\Delta^{G,i-1} \mathbf{1}_{G,i}}{\mathbf{1}_{G,i}' \Delta^{G,i-1} \mathbf{1}_{G,i}}.$$

Ce résultat correspond à une optimisation faible au sens suivant. Le poids w_i donné par (2.6) satisfait $E(w_i) = 1$ et de surcroît $E(w_i | i \in \Omega^B) = 1 / \pi_i^B$ où π_i^B est la probabilité d'inclusion de l'unité i dans Ω^B , qu'il est généralement difficile, voire impossible, de calculer en pratique. Notons que l'estimateur de Horvitz-Thompson est caractérisé par $\text{Var}(w_i | i \in \Omega^B) = 0$. L'optimisation faible obtenue ici revient à minimiser $\text{Var}(w_i | i \in \Omega^B)$ sur toutes les matrices de liens normalisées possibles Θ^{GB} , ou, de façon équivalente Θ^{AB} . Cette variance est strictement positive dans les cas où l'unité i de U^B peut recevoir plus qu'un seul poids pour divers échantillons s^d . En outre, si nous utilisons (6.3), le multiplicateur λ_i^t semble être la variance du poids w_i et est, par conséquent, toujours strictement positif (sauf, cas que nous excluons, quand l'unité i est sélectionnée avec un poids égal à 1).

6.4 Forte optimabilité indépendante de \mathbf{Y}

L'optimabilité faible est une condition nécessaire à l'optimabilité forte indépendante du vecteur \mathbf{Y} d'une variable d'intérêt. Elle donne la forme nécessaire des vecteurs $\Theta^{GB, \text{opt}, i}$ dans (6.4). Pour obtenir les conditions suffisantes pour une forte optimabilité indépendante de \mathbf{Y} , nous retournons aux équations (6.4). Ces dernières doivent être satisfaites pour tous les vecteurs \mathbf{Y} et doivent par conséquent être satisfaites pour une variable d'intérêt particulière, telle que $y_i = 1$ pour une unité i de U^B , $y_{i'} = 0$ pour une autre unité i' de U^B , et $y_{i''} = 0$ pour toutes les autres unités i'' de U^B ($i'' \neq i, i'$). Dans ce cas, nous pourrions avoir les relations suivantes pour tout i et i' :

$$(6.6) \quad \Delta^{G,i} \Theta^{GB, \text{opt}, i} + \Delta^{G,i'} \Theta^{GB, \text{opt}, i'} = \lambda_i^t \mathbf{1}_{G,i} \quad \Delta^{G,i'} \Theta^{GB, \text{opt}, i'} = \lambda_{i'}^t \mathbf{1}_{G,i'}$$

Comme nous devons nécessairement avoir une optimabilité faible, nous avons $\Delta^{G,i} \Theta^{GB, \text{opt}, i} = \lambda_i^t \mathbf{1}_{G,i}$. Partant de la première ligne de (6.6), nous obtenons alors

$$(6.7) \quad \Delta^{G,i'} \Theta^{GB, \text{opt}, i'} = (\lambda_i^t - \lambda_{i'}^t) \mathbf{1}_{G,i'}$$

En multipliant les deux membres de (6.7) par $\Theta^{GB, \text{opt}, i'}$ nous obtenons

$$\Theta^{GB, \text{opt}, i'} \Delta^{G,i'} \Theta^{GB, \text{opt}, i'} = \Phi^{i'} \Theta^{GB, \text{opt}, i'} \mathbf{1}_{G,i'}$$

$$= \Phi^{i'}$$

puisque $\Theta^{GB, \text{opt}, i} \mathbf{1}_{G,i} = 1$. Soit Φ la matrice contenant les éléments $\Phi^{i'}$ hors de la diagonale et $\Phi^{ii} = \lambda_i^t$ sur la diagonale. En utilisant de nouveau (6.2), nous pouvons

(3.2) par rapport à la matrice de liens normalisée Θ_{AB} . Par la factorisation présentée à la section 6.1, nous obtenons

$$\begin{aligned} \text{Var}(Y_B) &= Y' \Theta_{AB} \Delta^A \Theta_{AB} Y \\ &= Y' \Theta_{AB} \Theta_{AG} \Delta^A \Theta_{AG} \Theta_{GB} Y \\ &= Y' \Theta_{GB} \Delta^G \Theta_{GB} Y \end{aligned} \quad (6.2)$$

où $\Delta^G = \Theta_{AG} \Delta^A \Theta_{AG}$. Pour toute matrice de liens normalisée Θ_{AB} , la factorisation présentée à la section 6.1 produit systématiquement le même premier facteur Θ_{AG} . Par conséquent, si nous recherchons une matrice optimale $\Theta_{AB, \text{opt}}$ qui minimise la variance (3.2), il suffit d'optimiser le deuxième facteur Θ_{GB} . Nous aimerions aussi que la matrice optimale produise des estimations sans biais.

Soit U_G^i la sous-population de U_G contenant les N_G^i liens vers l'unité i de U_B . Notons que les sous-populations U_G^i sont disjointes. Donc, sans perte de généralité, nous pouvons classer les liens allant de U^A à U_B de façon que, pour tout i , les liens vers l'unité i dans U_B soient indexés de la matrice Θ_{GB} , $i = 1, \dots, N_B^i$. Par construction, le vecteur $\Theta_{GB, i}$ ne contient que des éléments non nuls pour les N_G^i liens vers l'unité i de U_B . Donc, si nous représentons par $\Theta_{GB, i}$ un vecteur colonne de taille N_G^i contenant les éléments non nuls de $\Theta_{GB, i}$, nous obtenons

$$\tilde{\Theta}_{GB, i} = \begin{bmatrix} 0 \\ \Theta_{GB, i} \\ 0 \end{bmatrix}$$

De même, soit $I_{G, i}$ le vecteur colonne de taille N_G^i contenant des valeurs 1 pour N_G^i éléments, et des valeurs 0 ailleurs. Si nous représentons par $I_{G, i}$ un vecteur colonne de taille N_G^i contenant les valeurs 1, nous obtenons

$$I_{G, i} = \begin{bmatrix} 0 \\ I_{G, i} \\ 0 \end{bmatrix}$$

Afin que l'application de la MGPP pour passer de U_G^i à U_B soit sans biais, il faut que nous ayons $\tilde{\Theta}_{GB, i} I_{G, i} = I_{G, i}$ pour toute i , ou de façon équivalente, $\tilde{\Theta}_{GB, i} I_{G, i} = I_{G, i}$. Ensemble, toutes ces considérations mènent au problème d'optimisation suivant :

Trouver une matrice $\Theta_{GB, \text{opt}} = \{\tilde{\Theta}_{GB, \text{opt}, 1}, \dots, \tilde{\Theta}_{GB, \text{opt}, N_B^i}\}$ satisfaisant $\tilde{\Theta}_{GB, \text{opt}, i} I_{G, i} = I_{G, i}$ pour tout $i = 1, \dots, N_B^i$, et minimisant la forme quadratique $\text{Var}(Y_B) = Y' \Theta_{GB} \Delta^G \Theta_{GB} Y$.

Ce problème n'est rien d'autre que la minimisation d'une forme quadratique positive sous des contraintes linéaires. Il s'agit d'un problème assez typique et simple à résoudre. Il est bien connu qu'il existe toujours une solution et qu'elle est unique si l'expression (6.2) est définie positive, ou que le

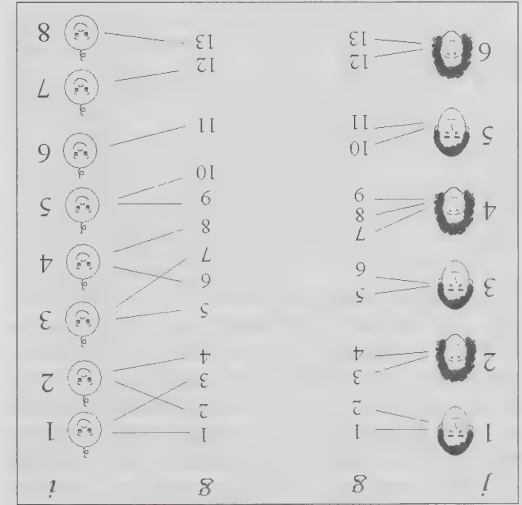


Figure 2. Résultat de la factorisation des populations parents-enfants.

Si nous considérons le graphe allant de U_G^i à U_B , nous pouvons construire la matrice de liens Θ_{GB} de taille $N_G \times N_B$ comme suit. Étant donné la définition de la population U_G , chaque unité g de U_G est liée à exactement une unité i de U_B . Notons que le sondage indirect dans ce contexte peut être considéré comme un échantillonnage de grappes (c'est-à-dire les unités i de U_B) à partir de leurs éléments (c'est-à-dire les unités g de U_G). Il peut également être considéré comme étant le cas « tous pour un à l'intérieur des grappes » présenté à la section 5.3. Soit $\Theta_{GB} = [\Theta_{GB, 1}, \dots, \Theta_{GB, N_B^i}]$ la matrice de liens normalisée obtenue à partir de Θ_{GB} . Nous avons $\text{diag}(I_{G, i}^c \Theta_{GB}) = \text{diag}(I_{G, i}^c \Theta_{AB})$, et, par conséquent, $\Theta_{GB} =$

$$\begin{aligned} \Theta_{AC} \Theta_{GB} &= \Theta_{AC} \Theta_{GB} \\ &= \Theta_{AC} [\text{diag}(I_{G, i}^c \Theta_{AB})]^{-1} \\ &= \Theta_{AB} [\text{diag}(I_{G, i}^c \Theta_{AB})]^{-1} \\ &= \Theta_{AB} \end{aligned} \quad (6.1)$$

Comme nous l'avons mentionné plus haut, le problème d'optimalité examiné ici consiste à minimiser la variance

5.5 Estimateur biaisé

Supposons que certaines colonnes de la matrice de liens Θ_{AB} ne contiennent que des zéros. Cela signifie que certaines unités de la population U^B ne sont associées à aucune unité de la population cible U^A . Rappelons que, pour que la matrice Θ_{AB} soit bien définie, il faut que $\text{diag}(\mathbf{1}^A \Theta_{AB})^{-1}$ existe. Comme nous le verrons, le cas qui nous occupe ne satisfait pas cette condition, ce qui mène à un estimateur biaisé du total Y^B .

De façon plus formelle, supposons que chacune des N^B premières colonnes de la matrice de liens Θ_{AB} contient au moins un $\theta_{ji} > 0$, et supposons qu'elles forment la sous-matrice Θ_1 , différentes de celles de la section précédente. Supposons que les N^B autres colonnes de Θ_{AB} ont $\theta_{ji} = 0$ pour $j = 1, \dots, N^A$. Nous avons par conséquent $\Theta_{AB} = [\Theta_1, \mathbf{0}]$.

De cette définition, il découle directement que

$$[\text{diag}(\mathbf{1}^A \Theta_{AB})]^{-1} = [\text{diag}(\mathbf{1}^A \Theta_1, \mathbf{1}^A \mathbf{0})]^{-1} = \begin{bmatrix} \text{diag}(\mathbf{1}^A \Theta_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^{-1} \quad (5.6)$$

Puisque cette matrice est singulière, $[\text{diag}(\mathbf{1}^A \Theta_{AB})]^{-1}$ n'existe pas. Il serait peut-être possible d'utiliser une *inverse généralisée* comme solution de ce problème. Rappelons que, pour une matrice carrée donnée \mathbf{A} , la matrice \mathbf{A}^- est une inverse généralisée de \mathbf{A} à condition que $\mathbf{A} \mathbf{A}^- \mathbf{A} = \mathbf{A}$ (Searle 1971). Une inverse généralisée possible de (5.6) est

$$[\text{diag}(\mathbf{1}^A \Theta_{AB})]^{-1} = \begin{bmatrix} \text{diag}(\mathbf{1}^A \Theta_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}^{-1} \quad (5.7)$$

Avec cette inverse généralisée, nous avons la matrice de liens normalisée suivante $\Theta_- = \Theta_{AB} [\text{diag}(\mathbf{1}^A \Theta_{AB})]^{-1} = [\Theta_1, \mathbf{0}]$. Partant de l'équation (2.6), nous pouvons obtenir le vecteur colonne \mathbf{W}_- de poids :

$$\mathbf{W}_- = \begin{bmatrix} \mathbf{0} \\ \mathbf{0}^T \mathbf{1}^A \Pi^{-1} \mathbf{1}^A \end{bmatrix} \quad (5.8)$$

Comme le montre l'expression (5.8), les poids sont nuls pour les unités i de la population cible U^B pour lesquelles Θ_{AB} contient $\theta_{ji} = 0$ pour $j = 1, \dots, N^A$. Partant de (2.4) $\mathbf{Y}_-^B = \mathbf{1}^A \mathbf{T}^A \Pi^{-1} \mathbf{1}^A \Theta_1 \mathbf{Y}_1$ où $\mathbf{Y}_1 = \{Y_1^1, \dots, Y_1^{N^B}\}$ est le sous-vecteur construit d'après les N^B premiers éléments de \mathbf{Y} . Puisqu'en général, $E(\mathbf{Y}_-^B) = \mathbf{1}^A \mathbf{Y}_1 \neq \mathbf{1}^A \mathbf{Y} = Y^B$, cet estimateur est biaisé pour le total Y^B .

6. Optimalité

L'optimalité est un aspect important de la MGPP. Comme nous l'avons montré au résultat 3, l'estimateur Y^B

obtenu par cette méthode fournit des estimations sans biais à condition que la matrice Θ_{AB} soit une matrice de liens normalisée. Étant donné que la variance (3.2) de cet estimateur dépend de cette matrice, il devrait exister au moins une matrice $\Theta_{AB, \text{opt}}$ telle que la variance de l'estimateur Y^B soit minimale. Autrement dit, nous aimerions trouver les valeurs que les éléments θ_{ji}^B plus grands que 0 devraient prendre pour obtenir l'estimateur de Y^B le plus précis.

Kalton et Brick (1995) ont été les premiers à examiner ce problème d'optimalité. Ils ont obtenu des résultats pour la situation simplifiée où $N^A = 2$ et où s^A est obtenu par échantillonnage avec probabilité égale. Ils ont conclu qu'il fallait utiliser $\theta_{AB, \text{opt}}^B = 1$ lorsque $\theta_{AB}^B > 0$ et $\theta_{AB, \text{opt}}^B = 0$ lorsque $\theta_{AB}^B = 0$. Lavallée (2002) et Lavallée et Caron (2001) ont obtenu des résultats du même genre par des simulations. Dans cette section, nous présentons de nouveaux résultats sur l'optimalité de la MGPP.

6.1 Factorisation

La factorisation est le problème inverse de la transitivity. Elle consiste à trouver une population U^C et des matrices de liens normalisées Θ_{AC} et Θ_{CB} telles que $\Theta_{AB} = \Theta_{AC} \Theta_{CB}$. Cet exercice simplifie considérablement la recherche d'une matrice de liens normalisée optimale $\Theta_{AB, \text{opt}}$.

Considérons que la population U^C est formée de grappes et que la factorisation est réalisée dans les contextes « un pour tous (à l'intérieur des grappes) » (de U^A à U^C) et « tous pour un (à l'intérieur des grappes) » (de U^C à U^B) présentés aux sections 5.2 et 5.3. Nous pouvons décrire cette situation de façon très générale comme il suit. Soit une population U^C contenant autant d'unités qu'il y a de liens partant des unités j de U^A . La taille de la population N^C est alors donnée par le nombre d'éléments $\theta_{AB, ji}^C$ de Θ_{AB} dont la valeur est supérieure à 0. Chaque unité g de U^C peut être conceptualisée comme étant l'extrémité d'une « flèche » partant d'une unité j de U^A . Partant de ce graphique, il n'existe qu'une seule matrice de liens Θ_{AC} de taille $N^A \times N^C$ assurant l'absence de biais, à savoir $\Theta_{AC} = [\theta_{AC, jg}^C]$, où $\theta_{AC, jg}^C = 1$ s'il existe un lien (ou une « flèche ») partant de l'unité j de U^A vers l'unité g de U^C , et $\theta_{AC, jg}^C = 0$ autrement. Notons que, par construction, chaque unité g de U^C est liée, au plus, à une unité j de U^A et, donc, que $\Theta_{AC} = \Theta_{AC}$. Cela correspond à la situation « un pour tous dans les grappes » présentée à la section 5.2. Le sondage indirect de U^A à U^C est en fait un sondage en grappes type et fait aboutir la MGPP à l'estimateur d'Horvitz-Thompson habituel (voir Lavallée 2002). Dans le cas de l'exemple des parents et des enfants, le résultat de cette factorisation serait donné par la figure 2.

Comme premier résultat, nous avons $\Theta^{AB} = \mathbf{I}$. Par conséquent, le vecteur de poids (2.6) est donné par $\mathbf{W}' = (t_1^N / \pi_1^N, \dots, t_N^N / \pi_N^N)$ et nous avons aussi $\mathbf{Z} = \Theta^{AB} \mathbf{Y} = \mathbf{Y}$. Donc, l'estimateur $\hat{\mathcal{Y}}^B$ donné par (2.5) n'est autre que l'estimateur d'Horvitz-Thompson $\hat{\mathcal{Y}}^B = \mathbf{I}' \mathbf{T}^{-1} \Pi^{-1} \mathbf{Y}$.

5.2 Un pour tous (à l'intérieur des grappes)

Considérons le cas où la population U^B est divisée en Γ grappes γ , chacune de taille N_γ^B . Ces grappes sont telles que chaque grappe γ de U^B est associée à exactement une unité j de U^A . Par conséquent, nous pouvons utiliser la lettre γ pour les unités j de U^A ainsi que pour les grappes de U^B . Notons aussi que $\Gamma = N^A$.

Cette situation correspond à une matrice de liens \mathbf{L}^{AB} de forme diagonale par blocs où chaque sous-matrice ne contient qu'une seule ligne. Soit le vecteur ligne $\mathbf{1}^{B_j}$ de taille N^{B_j} et contenant uniquement des 1. La matrice de liens \mathbf{L}^{AB} est alors définie comme étant

$$(I.5) \quad \begin{bmatrix} {}^{1g}I & 0 & \cdots & 0 \\ & \ddots & & \vdots \\ 0 & {}^{lg}I & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & {}^{lg}I \end{bmatrix} = {}^{gV}\Theta$$

Nous pouvons aussi écrire $\Theta_{AB} = \text{diag}(\{\Gamma^{B_1}, \dots, \Gamma^{B_r}\})$. En utilisant cette expression, nous avons $\text{diag}(\Gamma^A \Theta_{AB}) = \text{diag}(\Gamma^A \Gamma^{B_1}, \dots, \Gamma^A \Gamma^{B_r})$ et donc $\Theta_{AB} = \Theta^{AB}$. À partir de l'équation (2.6), nous obtenons le

vecteur colonne $\mathbf{W}' = (t_1^A, \pi_1^A, \dots, t_{B-1}^A, \pi_{B-1}^A)$ de poids \mathbf{W}' . Comme nous pouvons le voir, les éléments du vecteur colonne \mathbf{W} ont les valeurs t_i^A / π_i^A de U_B . Partant de (2.4), nous obtenons

$$(5.2) \quad {}^{\lambda}A_{\mu}^{\nu} \frac{{}^{\lambda}A_{\rho}^{\sigma}}{{}^{\lambda}A_{\tau}^{\eta}} \sum_{\gamma}^{\lambda} = {}^{\lambda}A_{\gamma}^{\eta} \sum_{\gamma}^{\lambda}$$

5.3 Tous pour un (à l'intérieur des grappes)

Considérons le cas où la population U^A est divisée en Γ grappes γ_i , chacune de taille N_i^A . Ces grappes sont telles que chaque grappe γ_i de U^A est associée à exactement une unité i de U^B . Par conséquent, nous pouvons utiliser la lettre γ pour les grappes de U^A , ainsi que les unités i de U^B . Notons aussi que $\Gamma = N_B$.

Cette situation correspond à une matrice de liens Θ^{AB} de forme diagonale par blocs où chaque sous-matrice ne contient qu'une seule colonne. Soit le vecteur colonne $\mathbf{1}^{A'}$ de taille $N^{A'}$ et contenant uniquement des 1. La matrice de liens Θ^{AB} est alors définie comme étant

Deville et Lavallée : Sondage indirect : Les fondements de la méthode généralisée du partage des poids

$$(53) \quad \Theta^{\mathcal{B}} = \begin{bmatrix} \mathcal{I} & 0 & \cdots & 0 \\ & \ddots & & \vdots \\ 0 & \mathcal{I} & 0 & 0 \\ & \vdots & & \ddots \\ 0 & \cdots & 0 & \mathcal{I} \end{bmatrix}$$

Nous pouvons aussi écrire $\Theta^{AB} = \text{diag}\{\lambda^{1A}, \dots, \lambda^{1A}, \lambda^{1B}, \dots, \lambda^{1B}\}$. En utilisant cette expression, nous avons $\Theta^{AB} = \lambda^{AB}$. D'après (2.6), nous obtenons le vecteur colonne des poids $\mathbf{W}' = (1/N^A \sum_{j=1}^N \lambda^{1A}_{ij} / \pi^A_j)$. Donc, les éléments γ (ou i) du vecteur colonne \mathbf{W} ont les valeurs moyennes $\sum_{j=1}^N \lambda^{1A}_{ij} / \pi^A_j$, $\gamma = 1, \dots, I$. Partant de (2.4), nous obtenons $\mathbf{Y}^B = \Sigma^{-1} \mathbf{X}^B / N^A \Sigma_{j=1}^N \lambda^{1A}_{ij} / \pi^A_j$.

5.4 Echantillonnage inefficace

Supposons que certaines lignes de la matrice de liens Θ_{AB} ne contiennent que des zéros. Cela signifie que certaines unités de la population U_A ne sont associées à aucune unité de la population U_B . Alors, si de telles unités sont sélectionnées dans l'échantillon s^A , elles ne permettront d'identifier aucune unité de U_B , ce qui peut être considéré comme inefficace du point de vue de l'échantillonnage. De façon plus formelle, supposons que chacune des N_A^j premières lignes de la matrice de liens Θ_{AB} contient au moins un $\theta_{ij}^j < 0$, et qu'elles forment la sous-matrice Θ_{AB}^1 . Supposons que les N_{AB} autres lignes de Θ_{AB} ont $\theta_{ij}^j = 0$ pour $i = 1, \dots, N_B$. Par conséquent, nous avons

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Comme premier résultat, nous obtenons

$$(5.4) \quad \begin{bmatrix} \mathbf{0} \\ \tilde{\Theta}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \Theta_1 [\text{diag}(\mathbf{I}'_A \Theta_1)]^{-1} \end{bmatrix} = \tilde{\Theta}^{AB}$$

où I^A est le vecteur colonne de 1 de taille N^A . Partant de l'équation (2), nous obtenons le vecteur colonne de poids $W = \Theta^A \begin{bmatrix} 0 \\ I^A \end{bmatrix} N^{-A}$. Soit $\Pi = \text{diag}\{I^A, \dots, I^A\}$ la matrice diagonale de taille $N^A \times N^A$ et de plus, soit $\mathbf{I} = \text{diag}\{I^A, \dots, I^A\}$ la matrice diagonale de taille $N^A \times N^A$. Nous obtenons alors

$$\begin{aligned} & \cdot^{\mathcal{F}^1} \mathbf{I}_1^{\mathcal{F}^1} \mathbf{H}^{\mathcal{F}^1} \mathbf{L}_1^{\mathcal{F}^1} \mathbf{Q} = \\ & \cdot^{\mathcal{F}^1} \mathbf{I}_1^{\mathcal{F}^1} \mathbf{H}^{\mathcal{F}^1} \mathbf{L}_1^{\mathcal{F}^1} [\mathbf{0}^{\mathcal{F}^1} \mathbf{Q}] = \mathbf{M} \end{aligned}$$

Comme le montre (5.5), les poids dépendent uniquement des probabilités de sélection π_i^j des unités de U^A qui ont au moins un $\theta_i^j > 0$ pour $i = 1, \dots, N_B$. À partir de (2.4), nous obtenons finalement $Y_B = I^A T^A \Pi^A \Theta^A Y$.

Supposons qu'on nous donne une fonction ϕ partant de l'ensemble de sous-ensembles de U^A vers l'ensemble de sous-ensembles de U^B . Comme f , supposons que ϕ satisfait la « propriété d'union » : $\phi(s_1^A \cup s_2^A) = \phi(s_1^A) \cup \phi(s_2^A)$, où s_1^A et s_2^A sont deux sous-ensembles de U^A .

Résultat 6 :

La fonction ϕ est déterminée sans équivoque par une matrice de liens zéro-un.

Démonstration :

Nous pouvons le démontrer comme il suit : prenons $s_j^A = \{j\}$ pour une unité j dans U^A . Alors, $\phi(s_j^A)$ est un ensemble dans U^B . Soit $l_{AB}^j = 1$ si l'unité j de U^B appartient à $\phi(s_j^A)$, et 0 autrement. En vertu de la propriété d'union, $\phi(s^A) = \bigcup_{j \in s^A} \phi(s_j^A)$ et l'ensemble de l_{AB}^j définit la matrice de liens zéro-un $L_{AB} = [l_{AB}^j]$ de taille $N^A \times N^B$, qui définit précisément la fonction ϕ .

Cela nous donne une relation d'équivalence entre les matrices de liens, associées à une propriété plus profonde.

Soit p^A un plan d'échantillonnage sur U^A (c'est-à-dire une loi de probabilité sur l'ensemble de sous-ensembles de U^A). La fonction f induit un plan d'échantillonnage sur U^B par $p^B(\Omega_B) = \sum_{s^A \in \Omega_B = f(s^A)} p^A(s^A)$. Comme le plan est induit par f , il ne dépend pas de la matrice de liens particulière Θ_{AB} définissant la fonction, mais est plutôt une

caractéristique de la classe d'équivalence par la voie de la matrice de liens zéro-un L_{AB} . Par conséquent, l'estimateur d'Horvitz-Thompson en U^B dépend uniquement de cette classe. Il y a donc un certain intérêt à choisir dans cette classe une matrice Θ_{AB} ayant, dans un certain sens, une caractéristique optimale (voir la section 6).

5. Matrices de liens spéciales

Comme le montre les sections précédentes, la matrice de liens Θ_{AB} dicte la forme de l'estimateur (2.4) donnée par la MGPP. À la présente section, nous décrivons certaines matrices de liens spéciales Θ_{AB} qui correspondent à des cas extrêmes. Il est probable que tous ces cas ne seront pas observés en pratique, mais ils illustrent l'effet de la matrice de liens sur l'estimateur (2.4).

5.1 Matrice identité

Supposons que la matrice de liens Θ_{AB} soit donnée par la matrice identité **I**. En pratique, cela signifie que la relation entre la population U^A et la population cible U^B est bijective. Naturellement, cela implique que $N^A = N^B = N$ et que la matrice identité **I** est de taille $N \times N$.

En remplaçant W^B par (3.4) dans l'équation (3.6), nous obtenons

$$W^C = \Theta^{BC} \text{diag}(W^B) \mathbf{1}^B. \quad (3.6)$$

Comme les poids $w_B^j = 0$ pour $j \notin \Omega_B$, nous avons $\mathbf{t}^B = (t_1^B, \dots, t_{N^B}^B)'$ et $t_i^B = 1$ si $i \in \Omega_B$, et 0 autrement. où $\Theta^{BC} = \Theta^{BC} [\text{diag}(\mathbf{t}^B \Theta^{BC})^{-1}]'$ et $\mathbf{T}^B = \text{diag}(\mathbf{t}^B)$ avec

$$W^C = \Theta^{BC} \text{diag}(\Theta^{AB} \mathbf{T}^A \Pi^{A-1} \mathbf{1}^A) \mathbf{1}^B \\ = \Theta^{BC} \Theta^{AB} \mathbf{T}^A \Pi^{A-1} \mathbf{1}^A. \quad (3.7)$$

Puisque $\Theta^{AB} \mathbf{1}^A = \Theta^{BC} \mathbf{1}^B = \mathbf{1}^C$, d'après le résultat 1,

la matrice $\Theta^{AB} \Theta^{BC}$ est une matrice de liens normalisée. Par conséquent, la MGPP est transitive, du moins dans un certain sens. Autrement dit, les poids W^C peuvent être obtenus en une seule étape en utilisant la matrice de liens normalisée $\Theta^{AB} \Theta^{BC}$ dans la MGPP. Maintenant, pour que la MGPP soit parfaitement transitive, les poids W^C donnés par (3.7) devraient être exactement les mêmes que les poids W^C donnés par (3.3). En comparant les équations (3.3) et (3.7), nous obtenons le résultat suivant :

Résultat 5 :

L'application de la MGPP de U^A à U^B , puis de U^B à U^C est transitive si et seulement si

$$\Theta^{AC} = \Theta^{AB} \Theta^{BC}. \quad (3.8)$$

Malheureusement, la condition (3.8) n'est pas vérifiée en général. En fait, il est relativement facile de produire des exemples où $\Theta^{AC} \neq \Theta^{AB} \Theta^{BC}$.

4. Une propriété structurelle de la MGPP

À la présente section, nous insistons sur le fait que, dans le cas du sondage indirect, le processus d'échantillonnage dépend uniquement des liens entre les deux populations U^A et U^B . Outre le fait d'être nuls ou non, les valeurs des θ_{AB}^{ij} proprement dites n'interfèrent pas avec le processus d'échantillonnage. Par ailleurs, les valeurs des θ_{AB}^{ij} jouent un rôle dans les poids (et donc l'estimateur) produits par la MGPP. Nous développons cette notion dans les paragraphes qui suivent.

L'échantillonnage indirect associé à chaque échantillon s^A dans U^A un échantillon Ω^B dans U^B , normalement $\Omega^B = \{j \in U^B \mid \exists j \in s^A \text{ et } \theta_{AB}^{ij} > 0\}$. Donc, une fonction $f: s^A \rightarrow \Omega^B$ qui établit la correspondance entre l'échantillon s^A et l'échantillon Ω^B est déterminée de façon unique par l'ensemble de couples (j, i) avec $\theta_{AB}^{ij} > 0$. Soit $t_{AB}^{ij} = 1$ si $\theta_{AB}^{ij} > 0$, et 0 autrement. Il s'agit des éléments de la matrice d'incidence du graphe reliant U^A à U^B .

3. Propriétés de la MGPP

3.1 Absence de biais

Comme l'a mentionné Ernst (1989), pour obtenir un estimateur sans biais, il suffit que $E(\mathbf{W}) = \mathbf{1}_B$. Par construction, puisque l'estimateur d'Horvitz-Thompson, cette condition est directement satisfaite et, par conséquent, la MGPP produit des estimations sans biais.

Partant de cette discussion, nous pouvons aussi obtenir le résultat suivant :

Résultat 3 :

Le vecteur de poids \mathbf{W} donné par (2.6) fournit des estimations sans biais si et seulement si la matrice Θ_{AB} est une matrice de liens normalisée.

Démonstration :

Partant de (2.6), nous avons

$$E(\mathbf{W}) = \Theta_{AB}^{-1} \mathbf{1}^A. \quad (3.1)$$

En utilisant le résultat 1, nous obtenons directement $E(\mathbf{W}) = \mathbf{1}_B$ et les estimations sont donc sans biais. Maintenant, supposons que $E(\mathbf{W}) = \mathbf{1}_B$. D'après (3.1), nous devons avoir $\Theta_{AB}^{-1} \mathbf{1}^A = \mathbf{1}_B$ et, par conséquent, Θ_{AB} est une matrice de liens normalisée.

3.2 Variance

Comme l'estimateur d'Horvitz-Thompson, nous obtenons directement le résultat suivant :

Résultat 4 :

La variance de \mathcal{Y}^B est donnée par

$$\begin{aligned} \text{Var}(\mathcal{Y}^B) &= \mathbf{Z}^A \Delta^A \mathbf{Z} \\ &= \mathbf{Y}^A \Delta^A \mathbf{Y} \end{aligned} \quad (3.2)$$

où $\Delta^A = [(\pi_{jj}^A - \pi_j^A \pi_j^A) / (\pi_j^A \pi_j^A)]_{N^A \times N^A}$ est une matrice définie non négative de taille $N^A \times N^A$ et où π_j^A est la probabilité de sélection conjointe des unités j et j' dans U^A , et où $\Delta^B = \Theta_{AB}^{-1} \Delta^A \Theta_{AB}$.

Pour une preuve de la variance de l'estimateur d'Horvitz-Thompson, voir Särndal, Swensson et Wretman (1992).

3.3 Transitivity

Supposons que nous voulions produire des estimations pour une population cible U^C que l'on ne peut atteindre que par l'entremise de la population U^B . Nous supposons que la population cible U^C contient N^C unités, chacune notée k . La correspondance entre les deux populations U^B et U^C peut être représentée par la matrice de liens $\Theta_{BC} = [\theta_{ik}^{BC}]$ de taille $N^B \times N^C$, où chaque élément $\theta_{ik}^{BC} \geq 0$. Autrement dit, l'unité i de U^B est reliée à l'unité k de U^C à condition que $\theta_{ik}^{BC} > 0$, sinon, il n'existe aucun lien entre les deux unités.

Considérons maintenant l'utilisation du sondage indirect en deux étapes. Pour chaque unité j sélectionnée dans s^A , nous identifions les unités i de U^B pour lesquelles la correspondance n'est pas nulle, c'est-à-dire pour lesquelles $\theta_{ij}^{AB} > 0$. Comme auparavant, nous avons $\theta_{ij}^{AB} = \{i \in U^B \mid \exists j \in s^A \text{ et } \theta_{ij}^{AB} > 0\}$. Pour chaque unité i de l'ensemble Ω_B , nous identifions alors les unités k de U^C pour lesquelles la correspondance n'est pas nulle, c'est-à-dire pour lesquelles $\theta_{ik}^{BC} > 0$. Nous avons alors l'ensemble $\Omega_C = \{k \in U^C \mid \exists i \in \Omega_B \text{ et } \theta_{ik}^{BC} > 0\}$. Parant de (2.6), nous obtenons le vecteur colonne \mathbf{W}^B de poids associés aux unités de la population U^B :

$$\mathbf{W}^C = \Theta_{AC}^{-1} \mathbf{T}^A \Pi^{-1} \mathbf{1}^A \quad (3.3)$$

où $\Theta_{AC} = \Theta_{AC} [\text{diag}(\mathbf{1}^A \Theta_{AC})]^{-1}$. Pour commencer, considérons le sondage indirect allant de U^A directement à la population cible U^C . Passer de la population U^A à U^B , puis à U^C revient à définir la matrice de liens $\Theta_{AC} = [\theta_{ik}^{AC}]$ de taille $N^A \times N^C$ par $\Theta_{AC} = \Theta_{AB} \Theta_{BC}$. Pour chaque unité j sélectionnée dans s^A , nous identifions les unités k de U^C pour lesquelles la correspondance n'est pas nulle, c'est-à-dire pour lesquelles $\theta_{jk}^{AC} > 0$, pour obtenir l'ensemble $\Omega_C = \{k \in U^C \mid \exists j \in s^A \text{ et } \theta_{jk}^{AC} > 0\}$. Nous mesurons la variable d'intérêt y_k^A à partir de la population cible U^C . En appliquant la MGPP, nous obtenons, d'après (2.6), les poids suivants :

Nous pouvons maintenant utiliser le sondage indirect par *transitivité*. Pour cela, nous sélectionnons un échantillon s^A de U^A et commençons par identifier l'ensemble Ω_B de U^B . À partir de cet ensemble Ω_B , nous identifions alors les unités de U^C qui y sont associées, afin de former l'ensemble $\Omega_C = \{k \in U^C \mid \exists i \in \Omega_B \text{ et } \theta_{ik}^{BC} > 0\}$. Une question importante est celle de savoir si, lorsqu'elle est appliquée dans le contexte du sondage indirect par transitivity, la MGPP est également transitive. Autrement dit, l'application de la MGPP de U^A à U^B , puis de U^B à U^C équivaut-elle à son application directe de U^A à U^C ?

Pour chaque unité i de l'ensemble Ω_B , nous avons alors un poids non nul w_i^B . Or, l'ensemble Ω_B peut être considéré comme un échantillon d'unités qui sont utilisées dans un processus de sondage indirect pour identifier l'ensemble Ω_C . Par similitude avec l'échantillonnage indirect allant de l'échantillon s^A à la population cible U^B , l'application de la MGPP dans le contexte du sondage indirect allant de l'ensemble Ω_B à la population cible U^C produit les poids suivants :

$$\mathbf{W}^B = \Theta_{AB}^{-1} \mathbf{T}^A \Pi^{-1} \mathbf{1}^A. \quad (3.4)$$

$$\mathbf{W}^C = \Theta_{BC}^{-1} \mathbf{T}^B \text{diag}(\mathbf{W}^B) \mathbf{1}^B \quad (3.5)$$

pour mesurer la variable d'intérêt y_i . Heureusement, dans la plupart des applications (par exemple, le cas parents-enfants susmentionné), le nombre de liens qui ont pour origine une unité donnée j de s^A est plus ou moins prévisible (par exemple, un parent a en général 1, 2 ou 3 enfants), ce qui facilite la détermination du nombre d'unités i de U_B qui, en dernière analyse, seront mesurées.

Nous supposons que pour toute unité j de s^A , il est possible d'obtenir les correspondances pour $i = 1, \dots, N_B$. Autrement dit, nous pouvons identifier tous les liens entre les deux populations par interview directe ou grâce à une source administrative pour toute unité j échantillonnée. En outre, pour toute unité i identifiée de U_B , nous supposons qu'il est possible d'obtenir les liens pour $j = 1, \dots, N^A$ (comme l'a mentionnée Lavallée (2002), il existe des cas où cette dernière contrainte est difficile à satisfaire en pratique. Si nous revenons à l'exemple des parents et des enfants, il pourrait être difficile pour un très jeune enfant, sélectionné par l'entremise de sa mère, de mentionner son père, si les parents sont divorcés. Afin de simplifier la discussion, nous supposons que ce genre de problème d'identification de liens est négligeable). Par conséquent, il n'est pas nécessaire de connaître les valeurs des liens entre les populations complètes U^A et U_B . En fait, nous ne devons connaître les liens (et, par conséquent, les valeurs de θ_{ij}^{AB}) que pour les lignes j de Θ^{AB} , où $j \in s^A$, ainsi que pour les colonnes i de Θ^{AB} , où $i \in \Omega_B$.

Supposons que nous voulions estimer le total Y_B de la population cible Y_B , où $Y_B = \sum_{i=1}^{N_B} y_i$. Nous pouvons aussi écrire $Y_B = \mathbf{1}^B Y$, où $\mathbf{1}^B$ est le vecteur colonne de 1 de taille N_B (notons que, pour simplifier, nous utilisons la notation $\mathbf{1}^B$ au lieu de $\mathbf{1}_{N_B}^B$). Maintenant, posons que $\theta_{ij}^{AB} = \sum_{f=1}^{N^A} \theta_{ijf}^{AB}$ et que $\theta_{ij}^{AB} / \theta_{ijf}^{AB} = \theta_{ij}^{AB+}$. Nous avons $\mathbf{1}^A \Theta^{AB} = \{\theta_{1j}^{AB+}, \dots, \theta_{N_B j}^{AB+}\}^T$. Nous définissons alors la *matrice de liens normalisée* $\Theta_{AB} = \Theta^{AB} (\mathbf{1}^A \Theta^{AB})^{-1}$, où $\text{diag}(\mathbf{v})$ est la matrice carrée obtenue en plaçant les éléments du vecteur \mathbf{v} sur la diagonale de 0 à 1. Notons que, pour que la matrice Θ_{AB} soit bien définie, il faut que $[\text{diag}(\mathbf{1}^A \Theta^{AB})]^{-1}$ existe, ce qui n'est le cas que si et autrement si $\theta_{ij}^{AB+} > 0$ pour tout $i = 1, \dots, N_B$. Dans l'exemple des parents et des enfants, cela signifie que chaque enfant doit être lié à au moins un parent.

Résultat 1 :

La matrice de liens Θ_{AB} est une matrice de liens normalisée si et seulement si

$$(2.1) \quad \Theta_{AB} \mathbf{1}^A = \mathbf{1}^B.$$

La preuve du résultat 1 découle directement de la définition d'une matrice de liens normalisée. Partant du résultat 1, nous obtenons directement le résultat 2 que l'on trouve aussi dans Deville (1998) :

Résultat 2 :

$$(2.2) \quad Y_B = \mathbf{1}^B Y$$

$$= \mathbf{1}^A \Theta_{AB} Y = \sum_{j=1}^{N_B} \sum_{i=1}^{N^A} \theta_{ij}^{AB+} y_i.$$

Soit le vecteur colonne $Z = \Theta_{AB} Y$ de taille N^A . En considérant chaque ligne de Z , nous définissons la variable $z_j = \sum_{i=1}^{N^A} \theta_{ij}^{AB+} y_i$ pour chaque unité j de la population U^A et nous la mesurons pour chaque unité $j \in s^A$.

Pour estimer Y_B , nous voulons utiliser les valeurs de y_i mesurées à partir de l'ensemble Ω_B . Pour cela, nous utiliserons un estimateur de la forme :

$$(2.3) \quad \hat{Y}_B = \sum_{j=1}^{N_B} w_j y_j$$

où w_j est le poids d'estimation de l'unité j de Ω_B , avec $w_j = 0$ pour $j \notin \Omega_B$. Soit $\mathbf{W}^T = \{w_1, \dots, w_{N_B}\}$. L'estimateur (2.3) peut être réécrit sous la forme

$$(2.4) \quad \hat{Y}_B = \mathbf{W}^T Y.$$

Habituellement, pour obtenir une estimation sans biais de Y_B , il suffit d'utiliser comme poids l'inverse de la probabilité de sélection π_B^i de l'unité i . Comme le mentionne Lavallée (1995) et Lavallée (2002), dans le cas du sondage indirect, il peut être difficile, voire impossible, de calculer cette probabilité. Il propose alors de recourir à la MGPP, qui est définie comme il suit.

Soit $\pi^A = \{\pi_1^A, \dots, \pi_{N^A}^A\}$ et soit $\Pi^A = \text{diag}(\pi^A)$ la matrice diagonale de taille $N^A \times N^A$ contenant les probabilités de sélection utilisées pour le tirage de l'échantillon s^A . Similairement, soit $\pi^B = \{\pi_1^B, \dots, \pi_{N_B}^B\}$ ou $\pi^B = \mathbf{1}^B$ si $j \in s^A$, et 0 autrement. Soit $\mathbf{T}^A = \text{diag}(\pi^A)$, la matrice diagonale de taille $N^A \times N^A$ contenant les variables indicatrices π_j^A . En partant de $Y_B = \mathbf{1}^A \Theta_{AB} Y$, nous pouvons former directement l'estimateur d'Horvitz-Thompson en fonction du vecteur Z :

$$(2.5) \quad \hat{Y}_B = \mathbf{1}^A \mathbf{T}^A \Pi^A \mathbf{Z}.$$

Puisque $Z = \Theta_{AB} Y$, nous avons $\hat{Y}_B = \mathbf{1}^A \mathbf{T}^A \Pi^A \Theta_{AB} Y$ et nous pouvons donc définir le vecteur colonne \mathbf{W} de poids :

$$(2.6) \quad \mathbf{W} = \Theta_{AB}^T \mathbf{T}^A \Pi^A \mathbf{1}^A.$$

Le vecteur \mathbf{W} est de taille N_B et, pour chaque $i = 1, \dots, N_B$, nous avons $w_i = \sum_{j=1}^{N^A} \theta_{ij}^{AB+} / \pi_j^A$. Les poids w_i de ce vecteur sont obtenus par la MGPP, comme le décrit Lavallée (2002).

les deux populations U^A et U^B peut être représentée par une *matrice de liens* $\Theta_{AB} = [\theta_{ij}^{AB}]$ de taille $N^A \times N^B$, où chaque élément est $\theta_{ij}^{AB} \geq 0$. Autrement dit, l'unité j de U^A est reliée à l'unité i de U^B à condition que $\theta_{ij}^{AB} > 0$; sinon, il n'existe aucun lien entre les deux unités. Dans le cas de l'exemple susmentionné, la matrice de liens est donnée par

$$\Theta_{AB} = \begin{bmatrix} \theta_{11}^{AB} & \theta_{12}^{AB} & \theta_{13}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \theta_{21}^{AB} & \theta_{22}^{AB} & \theta_{23}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_{31}^{AB} & \theta_{32}^{AB} & \theta_{33}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_{41}^{AB} & \theta_{42}^{AB} & \theta_{43}^{AB} & \theta_{44}^{AB} & \theta_{45}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{51}^{AB} & \theta_{52}^{AB} & \theta_{53}^{AB} & \theta_{54}^{AB} & \theta_{55}^{AB} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{61}^{AB} & \theta_{62}^{AB} & \theta_{63}^{AB} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

En sondage indirect, l'obtention de la *matrice de liens* $\Theta_{AB} = [\theta_{ij}^{AB}]$ est une question cruciale. Dans le cas où deux

unités $j \in U^A$ et $i \in U^B$ ne sont pas liées, nous fixons simplement que $\theta_{ij}^{AB} = 0$. Lorsqu'il existe un lien entre deux unités j et i , le choix de la valeur de $\theta_{ij}^{AB} > 0$ est important. Comme nous le verrons, il influe sur la précision des estimations émanant du sondage indirect. Dans plusieurs applications, les valeurs de θ_{AB}^{ij} pour les unités liées sont simplement fixées à 1. Naturellement, elles pourraient être choisies différentes de 1. Lavallée et Caron (2001) discutent l'utilisation des poids de couplage obtenus à partir d'un processus de couplage d'enregistrements entre U^A et U^B pour attribuer des valeurs aux éléments θ_{AB}^{ij} . Les poids de couplage sont proportionnels à la probabilité que deux unités $j \in U^A$ et $i \in U^B$ soient liées. Puisque le choix de $\theta_{AB}^{ij} > 0$ pour les deux unités liées j et i peut influencer la précision des estimations, il est naturel de rechercher les valeurs de θ_{AB}^{ij} qui minimiseront la variance des estimations. Ce problème d'optimisation est examiné à la section 6 de l'article.

Dans le sondage indirect, nous sélectionnons l'échantillon s^A de N^A unités à partir de U^A selon un certain plan d'échantillonnage. Soit π_j^A la probabilité de sélection de l'unité j . Nous supposons que $\pi_j^A > 0$ pour tout $j \in U^A$. Pour chaque unité j sélectionnée dans s^A , nous identifions les unités i de U^B pour lesquelles la correspondance n'est pas nulle, c'est-à-dire pour lesquelles $\theta_{AB}^{ji} > 0$. Soit Ω_B^j l'ensemble des N^B unités de U^B identifiées par les unités $j \in s^A$, c'est-à-dire $\Omega_B^j = \{i \in U^B \mid \exists j \in s^A \text{ et } \theta_{AB}^{ji} > 0\}$. Pour chaque unité i de l'ensemble Ω_B^j , nous mesurons une variable d'intérêt y_i à partir de la population cible de U^B . Soit $\mathbf{Y} = \{y_1, \dots, y_{N^B}\}$ le vecteur colonne de cette variable d'intérêt. D'un point de vue pratique, il est important de mentionner que, bien que la taille d'échantillon n^A soit habituellement déterminée d'avance, le nombre d'unités n^B est difficile à contrôler, car il dépend de l'échantillon sélectionné s^A et de la matrice de liens Θ_{AB} . Par conséquent, il s'avère difficile en général d'établir un budget

optimum dans un sens fort et indépendants de la variable d'intérêt.

Comme nous l'avons mentionné dans l'introduction, le sondage indirect consiste à sélectionner un échantillon s^A dans une population U^A afin de produire une estimation pour une population cible U^B , en s'appuyant pour cela sur la correspondance qui existe entre les deux populations. Par exemple, supposons que nous voulions produire des estimations pour une population d'enfants (unités de collecte), mais que nous ne disposons d'une base de sondage que pour les parents. La population cible U^B est celle des enfants, mais nous devons sélectionner un échantillon de parents avant de pouvoir interviewer les enfants. Cette situation est illustrée à la figure 1.

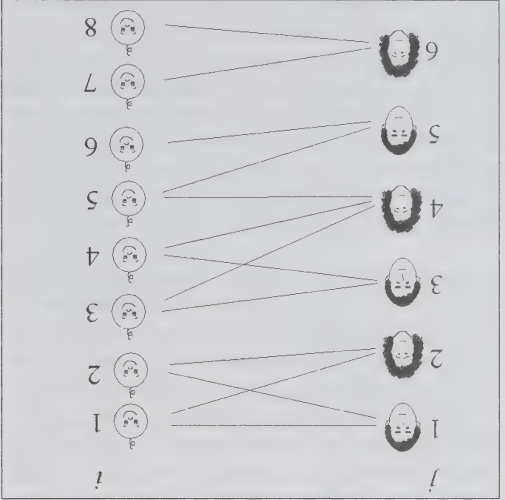


Figure 1. Population U^A de parents et population U^B d'enfants avec les liens entre les deux.

Soit U^A une population de N^A unités, où chaque unité est notée j . De même, soit U^B la population cible de N^B unités, où chaque unité est notée i . La correspondance entre

Sondage indirect : Les fondements de la méthode généralisée du partage des poids

Jean-Claude Deville et Pierre Lavalée¹

Résumé

Lorsqu'on veut sélectionner un échantillon, il arrive qu'au lieu de disposer d'une base de sondage contenant les unités de collecte souhaitées, on ait accès à une base de sondage contenant des unités liées d'une certaine façon à la liste d'unités de collecte. On peut alors envisager de sélectionner un échantillon dans la base de sondage disponible afin de produire une estimation pour la population cible souhaitée en s'appuyant sur les liens qui existent entre les deux. On donne à cette approche le nom de *sondage indirect*.

L'estimation des caractéristiques de la population cible étudiée par sondage indirect peut poser un défi de taille, en particulier si les liens entre les unités des deux populations ne sont pas bijectifs. Le problème vient surtout de la difficulté à associer une probabilité de sélection, ou un poids d'estimation, aux unités étudiées de la population cible. La méthode généralisée du partage des poids (MGPP) a été mise au point par Lavalée (1995) et Lavalée (2002) afin de résoudre ce genre de problème d'estimation. La MGPP fournit un poids d'estimation pour chaque unité enquêtée de la population cible.

Le présent article débute par une description du sondage indirect, qui constitue le fondement de la MGPP. En deuxième lieu, nous donnons un aperçu de la MGPP dans lequel nous la formons dans un cadre théorique en utilisant la notation matricielle. En troisième lieu, nous présentons certaines propriétés de la MGPP, comme l'absence de biais et la transitivity. En quatrième lieu, nous considérons le cas particulier où les liens entre les deux populations sont exprimés par des variables indicatrices. En cinquième lieu, nous étudions certains liens typiques spéciaux afin d'évaluer leur effet sur la MGPP. Enfin, nous examinons le problème de l'optimalité. Nous obtenons des poids optimaux dans un sens faible (pour des valeurs particulières de la variable d'intérêt), ainsi que les conditions dans lesquelles ces poids sont également optimaux au sens fort et indépendants de la variable d'intérêt.

Mots clés : Sondage indirect; méthode généralisée du partage des poids; absence de biais; poids optimaux.

1. Introduction

En vue de sélectionner les échantillons nécessaires pour les enquêtes sociales ou économiques, il est utile de disposer de bases de sondage, c'est-à-dire de listes d'unités, offrant un moyen d'atteindre les populations cibles souhaitées. Malheureusement, il arrive qu'au lieu de posséder une liste contenant les unités de collecte souhaitées, on dispose d'une liste d'unités reliée d'une certaine façon à celle des unités de collecte. On peut par conséquent partir de deux populations U^A et U^B liées l'une à l'autre, où l'on souhaite produire une estimation pour U^B . Malheureusement, on ne dispose d'une base de sondage que pour U^A . On peut alors envisager de sélectionner un échantillon s^A dans U^A afin de produire une estimation pour U^B en s'appuyant sur la correspondance qui existe entre les deux populations. On parle alors de *sondage indirect*.

L'estimation des caractéristiques d'une population cible U^B étudiée par sondage indirect peut poser un défi de taille, en particulier si les liens entre les unités des deux populations ne sont pas bijectifs. Le problème vient surtout de la difficulté à associer une probabilité de sélection, ou un poids d'estimation, aux unités de la population cible visées par le sondage. La méthode généralisée du partage des poids

Le but du présent article est de décrire le sondage indirect, c'est-à-dire les fondements de la MGPP, et d'obtenir, par la MGPP, des poids optimaux produisant des estimations sans biais dont la variance est minimale. Nous commencerons par décrire le sondage indirect, ainsi que la MGPP dans un cadre théorique qui fera appel, notamment, à la notation matricielle. L'utilisation de cette notation pour la MGPP a été présentée antérieurement par Deville (1998). Puis, nous utiliserons ce cadre théorique en vue d'énoncer certaines propriétés générales associées à la MGPP, dont l'absence de biais et la transitivity. Cette dernière consiste à passer de la population U^A à une population cible U^C par l'intermédiaire d'une population U^B . En troisième lieu, nous montrerons la correspondance entre la formulation

(MGPP) a été mise au point par Lavalée (1995) et Lavalée (2002), et également présentée dans Lavalée et Caron (2001), afin de résoudre ce genre de problème d'estimation. La MGPP fournit un poids d'estimation pour chaque unité étudiée de la population cible U^B . Fondamentalement, ce poids d'estimation correspond à une moyenne pondérée des poids de sondage des unités de l'échantillon s^A . La MGPP est une extension de la méthode de partage des poids décrite par Ernst (1989) dans le contexte des enquêtes longitudinales auprès des ménages.

ou que, lors des m (par exemple, 50) dernières étapes, la valeur de la fonction d'optimisation n'a pas varié. Enfin, calculer le vecteur \mathbf{k} (le vecteur de limites de strate finales) en fonction des valeurs du vecteur \mathbf{a} .

Bibliographie

- Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.
- Dalenius, T., et Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.
- Ekman, G. (1959). An approximation useful in univariate stratification. *Annals of Mathematical Statistics*, 30, 219-229.
- Glasser, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*, 30, 28-32.
- Gunning, P., et Horgan, J.M. (2004). Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques. *Techniques d'enquête*, 30, 177-185.
- Gunning, P., Horgan, J.M., et Yancey, W. (2004). Geometric stratification of accounting data. *J. de Contabilité y Administration*, 214, septembre-décembre.
- Hidiroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.
- Horgan, J.M. (2006). Stratification of skewed populations: A review. *Revue Internationale de Statistique*, 74(1): 67-76.
- Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5), 797-806.
- Lavallée, P., et Hidiroglou, M. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 35-45.
- Lechnicki, B., et Wiecekorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, 6, 287-306.
- Nelder, J.A., et Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7, 308-313.
- R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria: URL <http://www.R-project.org>.
- Rivest, L.-P. (2002). Une généralisation de l'algorithme de Lavallée aux fins de l'enquête annuelle sur les dépenses en capital du Bureau of the Census. *Techniques d'enquête*, 22, 65-75.
- Slania, J., et Krenzke, T. (1996). Utilisation de la méthode de Lavallée et Hidiroglou pour le calcul des limites de stratification aux fins de l'enquête annuelle sur les dépenses en capital du Bureau of the Census. *Techniques d'enquête*, 22, 65-75.
- Slania, J., et Krenzke, T. (1996). Utilisation de la méthode de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, 28, 207-214 (<http://www.mat.ulaval.ca/pages/lpr/>).
1. Trier la population en fonction des valeurs de la variable de stratification.
2. Choisir un vecteur initial \mathbf{a} , c'est-à-dire le vecteur de limites de strate initiales. Des nombres entiers aléatoires qui satisfont les contraintes peuvent être utilisés, mais la pratique révèle que de meilleurs résultats peuvent être obtenus en utilisant les limites de strate approximatives déterminées par une méthode de stratification approximative. Calculer la valeur de la fonction d'optimisation. Vérifier les contraintes; si elles ne sont pas satisfaites, les points initiaux doivent être modifiés.
3. Pour $r = 0, 1, \dots, R$ répéter l'étape suivante :
- a. Générer le point \mathbf{a}' en tirant une limite de strate a'_i puis en la modifiant comme il suit
- $$a'_i = a_i + f_i, \quad \text{for } k = 1, \dots, L - 1, k \neq i, \quad (11)$$
- où f est le nombre entier aléatoire, $f \in (-p; 1) \cup (1; p)$, p étant un nombre entier donné choisi d'après la taille de population (la valeur de p est d'autant plus élevée que la population est grande); habituellement, p devrait être compris entre 3 et 5.
- b. Calculer la valeur de la fonction d'optimisation.
- c. Si les contraintes sont satisfaites et que la valeur de la fonction d'optimisation sous le vecteur \mathbf{a}' est plus petite que celle obtenue sous le vecteur \mathbf{a} , accepter le nouveau vecteur, c'est-à-dire $\mathbf{a}^{r+1} = \mathbf{a}'$ (où \mathbf{a}^{r+1} est le vecteur de limites de strate dans une itération suivante); sinon, ne pas accepter le vecteur, c'est-à-dire $\mathbf{a}^{r+1} = \mathbf{a}$.
4. Finir l'algorithme si la règle d'arrêt est satisfaite, c'est-à-dire si $r = R$, où R est le nombre donné d'étapes

des échantillons provenant de certaines strates soit inférieure à deux ou supérieure à la taille de population de la strate.

Dans notre étude, l'approche par optimisation (au moyen des algorithmes LH et de recherche aléatoire) s'est avérée plus efficace que la stratification géométrique pour chaque population étudiée et chaque nombre de strates construites. Néanmoins, les limites de strate données par la stratification géométrique peuvent être considérées comme de bons paramètres initiaux pour l'approche par optimisation; par contre, elles ne devraient pas être regardées comme des limites de strate optimales ou efficaces. De surcroît, nos résultats montrent de façon concluante que la stratification géométrique est moins efficace que celle présentée par Lavallée et Hidiroglou (1988), résultat opposé à celui obtenu par Cumming et Horgan (2004) et par Horgan (2006). L'étude de ce problème doit se poursuivre sur des populations asymétriques réelles; les recherches portant sur des populations artificielles indiquent sans équivoque que l'algorithmes LH et l'approche par optimisation sont plus efficaces que la stratification géométrique.

À première vue, on pourrait s'étonner du fait que le gain d'efficacité réalisé en appliquant les approches LH et par optimisation comparativement à la stratification géométrique s'accroît lorsque le nombre de strates augmente. Toutefois, l'explication est simple. Le but de la stratification géométrique est d'égaliser les coefficients de variation de la variable de stratification dans les strates. Par conséquent, il diffère de celui de la stratification consistant à optimiser l'efficacité de l'estimation ou à minimiser la taille d'échantillon. Qui plus est, il n'est pas certain que la stratification optimale, la distribution de la variable de stratification/d'enquête soit uniforme dans les strates. Les deux ensembles de limites de strates (c'est-à-dire ceux fournis par les approches géométrique et par optimisation) ne sont pas nécessairement les mêmes; en fait, il est probable qu'ils diffèrent.

Notons que nous avons appliqué l'algorithmes de recherche aléatoire dans l'approche de stratification par optimisation. Or, l'algorithmes de Lavallée et Hidiroglou (1988) est également un représentant des approches par optimisation. Quand le but de la stratification est de minimiser la taille d'échantillon requise pour obtenir un niveau souhaité de précision, il est probable que les deux approches produisent des résultats semblables, comme cela a été le cas lors de notre étude. Néanmoins, l'algorithmes de recherche aléatoire peut être appliqué à n'importe quel problème de stratification (c'est-à-dire à toute fonction d'optimisation et ses contraintes), contrairement à l'algorithmes LH, qui n'est applicable qu'à la minimisation de la taille d'échantillon pour un niveau de précision donné. Il convient de souligner que l'algorithmes de recherche aléatoire fournit, comme méthode d'optimisation globale, des résultats aléatoires.

Notre but n'était pas, toutefois, de promouvoir l'un ou l'autre de ces deux algorithmes en montrant qu'ils sont plus efficaces que la stratification géométrique. Qui plus est, nous avons appliqué la méthode du simplexe de Nelder et Mead (1965) pour stratifier les populations (résultats non présentés ici); les résultats obtenus par cette méthode étaient fort semblables à ceux produits par les algorithmes LH et de recherche aléatoire. Chacune de ces méthodes présente certains inconvénients. Par exemple, des difficultés numériques peuvent survenir lors de l'utilisation de l'algorithmes LH (Slania et Krenzke 1996), tandis que la méthode de recherche aléatoire fournit des résultats aléatoires (Kozak 2004), la méthode de Nelder et Mead (1965) peut être inefficace si le nombre de strates et la taille de la population sont grands (Kozak 2004) et, en fait, l'obtention de points de stratification optimaux n'a été prouvée pour aucune de ces méthodes. Par conséquent, il reste encore à construire un algorithme de stratification produisant des résultats optimaux quelle que soit la situation (par exemple en ce qui concerne la taille de la population ou l'asymétrie de la variable), ainsi que des résultats non aléatoires. Notre objectif principal était de prouver que la stratification géométrique n'est pas optimale, mais que les points de stratification qu'elle produit peuvent être utiles comme paramètres initiaux dans d'autres approches de stratification.

Remerciements

Les auteurs remercient vivement les examinateurs et le rédacteur adjoint de *Techniques d'enquête* de leurs commentaires précieux, qui leur ont permis d'améliorer la première version du présent article.

Annexe

L'algorithmes qui suit a été proposé par Kozak (2004) et nous nous sommes bornés à adapter certains de ses détails au problème général de la stratification. Dans l'algorithmes, nous ne faisons pas référence au problème particulier de la stratification (autrement dit, nous ne définissons pas la fonction d'optimisation et ses contraintes), puisqu'il fonctionne pour les deux problèmes présentés dans l'article, ainsi que pour d'autres problèmes de stratification. Au besoin, nous faisons référence à la « fonction d'optimisation » (qui peut être la variance d'un estimateur étudié ou la taille d'un échantillon provenant d'une population) et aux « contraintes » (qui, selon la fonction d'optimisation, peuvent être les contraintes (5) et (6), ou les contraintes (5) combinées à la contrainte sur le niveau de précision de l'estimation); d'autres formes de la fonction d'optimisation et de ses contraintes peuvent sans aucun doute être prises en considération.

II découle des résultats que l'approche par optimisation est plus efficace que la stratification géométrique; cette observation a été faite pour chaque population et chaque nombre de strates. L'efficacité relative était systématiquement supérieure à 1,6. En outre, une conclusion intéressante se dégage de la comparaison de l'efficacité des stratifications géométrique et LH. Comme nous l'avons déjà mentionné, Gunning et Horgan (2004), ainsi que Horgan (2006) ont constaté que la stratification géométrique était plus efficace que l'algorithme LH. Par contre, dans notre étude, l'algorithme LH était systématiquement plus efficace que la stratification géométrique; constatation que nous avons également faite pour d'autres populations de taille et d'asymétrie différentes que nous avons générées (les résultats ne sont pas présentés ici). Néanmoins, nous constatons pas que l'algorithme LH est systématiquement plus efficace que la stratification géométrique. Il peut arriver que cette dernière donne de meilleurs résultats, comme Gunning et Horgan (2004) et Horgan (2006) l'ont observé lors de leurs études.

De la comparaison de l'algorithme LH à l'approche par optimisation, il découle que les deux méthodes donnent des points de stratification qui produisent des tailles d'échantillon semblables. Dans certains cas, la stratification LH est un peu meilleure et dans d'autres, un peu moins bonne, que l'approche par optimisation. Néanmoins, ces différences ne nous permettent pas de déclarer que l'une de ces deux approches est plus efficace que l'autre. En fait, elles ont toutes deux le même objectif (dans ce problème de stratification particulier) et diffèrent simplement en ce qui concerne l'algorithme utilisé pour atteindre cet objectif. Brevement, d'après nos résultats, nous concluons qu'en général, la stratification LH et l'approche par optimisation sont plus efficaces que la stratification géométrique.

5. Conclusion

La méthode de stratification fondée sur une progression géométrique proposée par Gunning et Horgan (2004) possède un avantage significatif; plus précisément, son algorithme est très simple à appliquer comparativement à la méthode de la fonction cumulative de la racine carrée des fréquences de Dalenius et Hodges (1959) et à d'autres méthodes de stratification. Toutefois, il s'agit d'une méthode approximative, si bien que les limites de strate qu'elle produit peuvent mener à des estimations de précision médiocre (ou nécessiter l'utilisation d'un échantillon de grande taille pour obtenir le niveau requis de précision). En outre, il est probable que les strates constituées ne satisfieront pas toutes aux contraintes (5); autrement dit, il se peut que certaines strates soient vides (de sorte qu'elles ne contiendront aucune unité de population) ou (et) que la taille

La présente section, nous comparons les trois approches de stratification, à savoir la stratification géométrique, l'algorithme LH et l'approche par optimisation selon la méthode de recherche aléatoire. Nous avons utilisé pour la présente étude les cinq mêmes populations qu'à la section précédente (voir tableau 1 et figure 1). Les efficacités relatives des approches comparées ont été évaluées au moyen de la formule

$$eff_{i,j} = \frac{n_j(cv)}{n_i(cv)}, \quad (10)$$

où i et j sont les indices des approches de stratification ($i, j = geom, optim, LH$), et $n_i(cv)$ et $n_j(cv)$ sont les tailles d'échantillon minimales requises pour obtenir un niveau souhaité de précision (cv) sous les i^e et j^e approches, respectivement. En suivant ces trois approches, nous avons stratifié chaque population en $L = 4, \dots, 7$ strates; le niveau fixé de précision était de 0,01 dans chaque cas. Les tailles minimales d'échantillon requises pour ce niveau de précision et les efficacités relatives (10) sont données au tableau 3.

Tableau 3
Tailles d'échantillon minimales requises pour obtenir une valeur égale à 0,01 pour le coefficient de variation de l'estimateur de la moyenne de population, sous la stratification géométrique (n_{geom}), l'approche par optimisation (n_{optim}) et l'algorithme LH (n_{LH}); et l'efficacité de la stratification géométrique relativement à l'approche par optimisation ($eff_{geom, optim}$), de la stratification géométrique relativement à l'algorithme LH ($eff_{geom, LH}$) et de l'algorithme LH relativement à l'approche par optimisation ($eff_{LH, optim}$)

Nombre de strates	L	n_{geom}	n_{optim}	n_{LH}	$eff_{geom, optim}$	$eff_{geom, LH}$	$eff_{LH, optim}$
4	4	805	496	496	1,63	1,63	1,00
5	5	613	344	344	1,78	1,78	1,00
6	6	460	252	252	1,83	1,83	1,00
7	7	357	192	192	1,86	1,86	1,00
4	4	483	248	259	1,94	1,86	1,04
5	5	329	154	163	2,14	2,02	1,06
6	6	224	113	117	1,98	1,92	1,03
7	7	180	83	83	2,17	2,17	1,00
4	4	782	410	411	1,91	1,90	1,00
5	5	601	303	304	1,98	1,98	1,00
6	6	422	241	242	2,04	2,05	1,00
7	7	496	195	195	2,11	2,16	1,00
4	4	839	409	409	2,05	2,05	1,00
5	5	650	301	301	2,15	2,15	1,00
6	6	552	240	242	2,30	2,28	1,01
7	7	— ¹	200	200	—	—	1,00
4	4	1 768	894	894	1,98	1,98	1,00
5	5	1 274	628	628	2,03	2,03	1,00
6	6	949	459	459	2,07	2,07	1,00
7	7	758	355	355	2,13	2,13	1,00

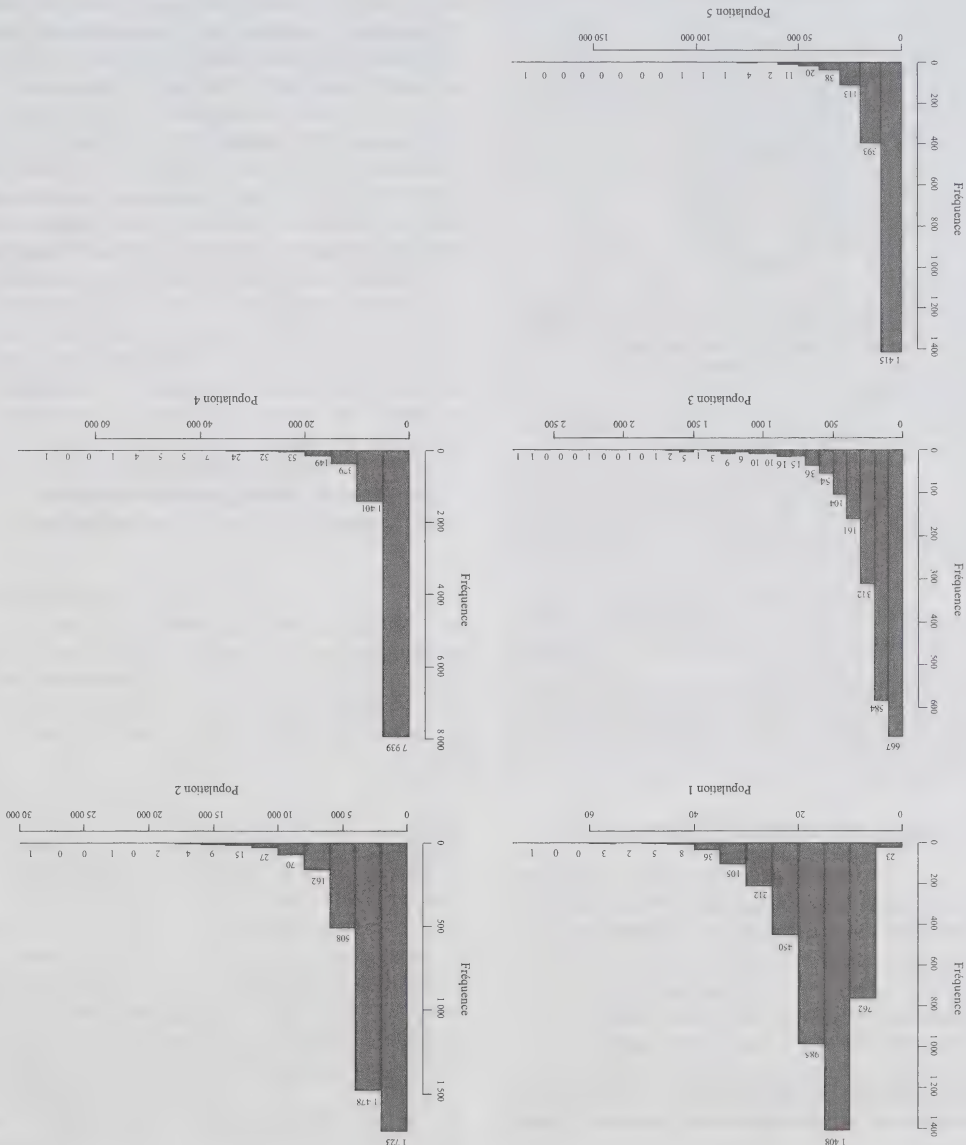
¹ L'obtention des limites de strate à posé des problèmes numériques (les tailles d'échantillon provenant de certaines strates étaient supérieures aux tailles de ces strates).

Dans chaque cas, l'approche par optimisation a été plus efficace que la stratification géométrique. L'efficacité n'était inférieure à 1,5 que pour deux combinaisons; pour les autres, elle variait entre 1,5 et 2. En général, le gain d'efficacité est d'autant plus important que le nombre de strates construites est grand.

Gunning et Horgan (2004), ainsi que Horgan (2006) ont comparé la stratification géométrique à l'algorithme de Lavallée et Hidiroglou (Lavallée et Hidiroglou 1988) et constaté que la première était généralement plus efficace. À

4. Comparaison numérique de l'efficacité des approches de stratification sous niveau de précision fixe de l'estimation

Figure 1. Histogrammes de la variable de stratification dans les populations artificielles étudiées.



limites de strate est meilleur que le précédent, il remplace ce dernier. L'annexe décrit en détail l'algorithme basé sur

l'article publié par Kozak (2004).

Le deuxième problème examiné dans le présent article est la construction de strates qui minimisent la taille de l'échantillon provenant d'une population sachant le niveau

de précision voulu de l'estimation (précision qui est donnée par la variance d'un estimateur de la moyenne ou du total de

population). L'algorithme de Lavallée-Hidiroglou (LH) (Lavallée et Hidiroglou 1988) peut être considéré comme une méthode d'optimisation particulière en vue de résoudre

ce problème précis de stratification; par contre, il n'est pas applicable à d'autres problèmes, par exemple celui consi-

déré plus haut. Pour des précisions sur l'algorithme, consulter l'article de Lavallée et Hidiroglou (1988). Outre

l'algorithme LH, nous avons appliqué la méthode de stratification géométrique et de recherche aléatoire pour

construire les strates.

Nous avons utilisé le langage et l'environnement R (R Development Core Team 2005) pour réaliser tous les

calculs de la présente étude.

3. Comparaison numérique de l'efficacité des approches de stratification sous taille d'échantillon fixe

À la présente section, nous comparons deux approches

de stratification, la stratification géométrique (geom) et l'approche par optimisation (optim), appliquées à un

problème de recherche des limites de strate qui minimisent la variance de l'estimateur considéré sachant une taille fixe

d'échantillon. Pour réaliser la comparaison, nous avons

généré cinq populations artificielles de tailles différentes (allant de 2 000 à 10 000). Les statistiques sommaires de ces

populations sont présentées au tableau 1; les histogrammes

des variables de stratification dans les populations sont

donnés à la figure 1. Dans chaque cas, la variable de

stratification était positivement asymétrique (le coefficient

d'asymétrie variait de 1,40 pour la première population à 5,02 pour la cinquième). Comme cela est généralement le

Tableau 1
Statistiques sommaires pour les populations artificielles étudiées

Population	Taille	Étendue	Asymétrie	Moyenne	Variance
1	4 000	3-22	1,40	16,11	45,8
2	4 000	243-28 578	2,66	2 823,95	$4,8 \times 10^6$
3	2 000	6-2 793	3,55	224,12	$6,0 \times 10^4$
4	10 000	62-74 398	4,20	3 616,41	$2,1 \times 10^7$
5	2 000	259-186 685	5,02	9 265,36	$1,1 \times 10^8$

Comme Gunning et Horgan (2004), pour comparer l'efficacité des deux approches, nous avons calculé l'effi-

cacité relative en appliquant la formule :

$$(8) \quad \text{eff}_{\text{geom, optim}} = \frac{V_{\text{geom}}(x_{\text{st}})}{V_{\text{optim}}(x_{\text{st}})}$$

où $V_{\text{geom}}(x_{\text{st}})$ et $V_{\text{optim}}(x_{\text{st}})$ sont les variances (1) sous les approches géométrique et par optimisation, respectivement. En outre, nous avons calculé les coefficients de variation de

l'estimateur de la moyenne de population sous les deux

approches :

$$(9) \quad \text{cv}_{\text{geom}} = \frac{\sqrt{V_{\text{geom}}(x_{\text{st}})}}{\bar{x}_{\text{st}}}, \text{cv}_{\text{optim}} = \frac{\sqrt{V_{\text{optim}}(x_{\text{st}})}}{\bar{x}_{\text{st}}}$$

Le tableau 2 contient les valeurs des efficacités relatives (8) et des coefficients de variation (9) pour chaque combinaison étudiée (population \times nombre de strates).

Tableau 2

Coefficients de variation de l'estimateur de la moyenne de population sous les approches de stratification géométrique (CV_{geom}) et par optimisation (CV_{optim}), et efficacité de la stratification géométrique comparativement à l'approche par optimisation ($\text{eff}_{\text{geom, optim}}$)

Nombre de strates

CV_{geom} CV_{optim} $\text{eff}_{\text{geom, optim}}$

Population 1

4 0,0086 0,0070 1,53

5 0,0070 0,0042 1,66

6 0,0057 0,0034 1,66

7 0,0051 0,0029 1,75

Population 2

4 0,0116 0,0084 1,37

5 0,0095 0,0065 1,47

6 0,0085 0,0051 1,66

7 0,0073 0,0042 1,72

Population 3

4 0,0235 0,0133 1,76

5 0,0174 0,0100 1,74

6 0,0146 0,0081 1,80

7 0,0129 0,0067 1,91

Population 4

4 0,0104 0,0063 1,64

5 0,0089 0,0047 1,88

6 0,0073 0,0038 1,93

7 0,0064 0,0032 2,00

Population 5

4 0,0235 0,0134 1,76

5 0,0185 0,0100 1,86

6 0,0161 0,0080 2,00

7 0,0134 0,0074 1,82

positivement asymétrique) que les méthodes de stratification ne comportant pas la construction d'une strate à tirage complet. La stratification géométrique ne comprend pas la création d'une telle strate (Gunning et Horgan 2004). Le but du présent article est de comparer l'efficacité de la stratification géométrique, proposée par Gunning et Horgan (2004) à celle de deux approches de stratification par optimisation (Lavallée et Hidiroglou 1988; Lednicki et Wiecezorkowski 2003; Kozak 2004) fondées sur l'utilisation de méthodes numériques d'optimisation.

2. Approches de stratification comparées

Supposons que nous souhaitons stratifier une population sur un vecteur $\mathbf{x} = (x_1, \dots, x_N)^T$ de dimension N connu dès le départ (c'est-à-dire avant le début de l'étude) des valeurs d'une variable de stratification X .

Dans le présent article, nous considérons deux problèmes de stratification. Le premier consiste à construire L strates

sachant la taille fixe d'échantillon n . Supposons que nous recherchions un vecteur de dimension $(L + 1)$ de limites de strate $\mathbf{k} = (k_0, \dots, k_L)^T$, $(k_0 < \dots < k_L, k_0$ étant la valeur minimale et k_L la valeur maximale de X) qui minimise la variance d'un estimateur de la moyenne de population de X sous échantillonnage stratifié avec échantillonnage aléatoire simple sans remise dans les strates (STS) et combiné à une approche avec strate à tirage complet. (Il convient de souligner que nous traitons la variable de stratification comme étant identique à la variable d'enquête correspondante.) La variance de \bar{x}_{st} est donnée par

$$V(\bar{x}_{st}) = \sum_{l=1}^L \left(\frac{N}{N_h} \right) \left(1 - \frac{N_h}{n_h} \right) \left(\frac{S_h^2}{S^2} \right), \quad (1)$$
$$\bar{x}_{st} = \sum_{l=1}^L \frac{N_h}{N} \bar{x}_h, \quad \bar{x}_h = \frac{1}{n_h} \sum_{k=1}^{n_h} x_{hk}, \quad (h = 1, \dots, L),$$

où n_h est la taille de l'échantillon provenant de la h^e strate, N_h est la taille de la h^e strate, S_h^2 est la variance de population de X restreinte à la h^e strate, \bar{x}_{st} est l'estimateur de la moyenne de population de X sous échantillonnage STS , \bar{x}_h est l'estimateur de la moyenne de population de X dans la h^e strate sous échantillonnage aléatoire simple sans remise (SI) et x_{hk} est la valeur de X pour la k^e unité d'échantillonnage de la h^e strate et $h = 1, \dots, L$. La répartition optimale de l'échantillon, qui s'obtient dans le cas de notre problème, par minimisation de la variance (1) sachant la taille d'échantillon n , est donnée par l'approche avec strate à tirage complet (Lednicki et Wiecezorkowski 2003) :

$$n_h = (n - n_L) \left(\frac{N_h S_h^2}{\sum_{h=1}^{L-1} N_h S_h^2} \right), \quad h = 1, \dots, L - 1, \quad (2)$$

L'approche géométrique de stratification a pour objectif de rendre égales les valeurs du coefficient de variation de X dans les L strates. Elle consiste simplement à appliquer la formule qui suit basée sur une progression géométrique (Gunning et Horgan 2004)

$$k_h = ar^h, \quad h = 0, \dots, L, \quad (3)$$

où $a = \min(X)$, $k_L = \max(X)$ et $r = (k_L/k_0)^{1/L}$. La formule (3) repose sur l'hypothèse selon laquelle X suit une loi uniforme dans chaque strate.

L'approche par optimisation appliquée à ce problème de stratification particulier s'inspire de l'optimisation numérique du problème suivant : minimiser

$$f(\mathbf{k}) = V(\bar{x}_{st}), \quad (4)$$

où $V(\bar{x}_{st})$ est la variance (1) sous la répartition optimale (2), sous les contraintes

$$N_h \geq 2 \text{ et } 2 \leq n_h \leq N_h \text{ pour } h = 1, \dots, L - 1, \quad (5)$$
$$\sum_{h=1}^L n_h = n - n_L. \quad (6)$$

Parfois, si l'on veut que le niveau de précision soit plus ou moins le même dans chaque strate, il est possible d'appliquer une méthode de « répartition avec puissance » (en anglais, *power allocation*) (Bankier 1988; Rivest 2002; Lednicki et Wiecezorkowski 2003):

$$n_h = \frac{(n - n_L)(N_h x_h^p)}{\sum_{h=1}^{L-1} (N_h x_h^p)}, \quad p \in (0, 1], \quad h = 1, \dots, L - 1. \quad (7)$$

L'approche par optimisation est plus difficile à appliquer que l'approche géométrique, en grande partie parce que cette dernière requiert un algorithme considérablement plus simple. Un choix doit être fait parmi les diverses méthodes d'optimisation disponibles. Lednicki et Wiecezorkowski (2003) ont utilisé la méthode du simplexe de Nelder et Mead (1965); cependant, il est également possible d'appliquer des méthodes plus efficaces, qui nécessitent souvent l'auto-application d'algorithmes (par exemple, Kozak 2004). Il convient de souligner que la stratification géométrique ne tient compte ni de la formule de la variance (1), ni de la répartition de l'échantillon (2), ni des contraintes (5). Or, il peut arriver que l'une des contraintes (5) ne soit pas satisfaisante. Par conséquent, la stratification géométrique est une méthode de stratification approximative. Dans la présente étude, nous avons appliqué l'algorithme proposé par Kozak (2004) pour stratifier plusieurs populations. Il s'agit d'un algorithme de recherche aléatoire adapté au problème de la stratification. Cet algorithme est simple, à chaque étape, une limite de strate est sélectionnée aléatoirement et modifiée aléatoirement. Si le nouvel ensemble de

Approche de la stratification par une méthode géométrique et par optimisation : Une comparaison de l'efficacité

Marcin Kozak et Med Ram Verma

Résumé

L'article donne une comparaison des approches de la stratification par une méthode géométrique, par optimisation et par la méthode de Lavallée et Hidiroglou (LH). L'approche géométrique est une approximation, tandis que les méthodes de Lavallée et Hidiroglou et l'optimisation sont des méthodes numériques, peuvent être considérées comme des méthodes de stratification optimales. L'algorithme de la stratification géométrique est très simple comparativement à ceux des deux autres approches, mais il ne tient pas compte de la construction d'une strate à tirage complet, qui est habituellement produite lorsque l'on stratifie une population positivement asymétrique. Dans le cas de la stratification par optimisation, on peut prendre en considération toute forme de la fonction d'optimisation et de ses contraintes. Une étude numérique comparative portant sur cinq populations artificielles positivement asymétriques a indiqué que, dans chaque cas étudié, l'approche par optimisation était plus efficace que la stratification géométrique. En outre, nous avons comparé les approches géométrique et par optimisation à l'algorithme LH. Cette comparaison a révélé que la méthode géométrique de stratification était moins efficace que l'algorithme LH, tandis que l'approche par optimisation était aussi efficace que cet algorithme. Néanmoins, les limites de strate déterminées par la stratification géométrique peuvent être considérées comme de bons points de départ pour l'approche par optimisation.

1. Introduction

Gunning et Horgan (2004) ont proposé un algorithme de stratification basé sur une régression géométrique. Par

sous-échantillonnage, nous appellerons cette technique « approche géométrique de stratification », « stratification géométrique » ou simplement « approche géométrique ». La stratification géométrique vise à produire des valeurs égales du coefficient de variation d'une variable de stratification dans les diverses strates, en émettant l'hypothèse que la variable suit une loi uniforme dans chaque strate. Gunning et Horgan (2004) ont montré que leur algorithme est nettement plus facile à appliquer et plus efficace que la méthode classique de la fonction cumulée de la racine carrée des fréquences (Dalenius et Hodges 1959) et que l'algorithme de Lavallée et Hidiroglou (LH) (Lavallée et Hidiroglou 1988). Horgan (2006) a comparé la stratification géométrique aux méthodes de Dalenius et Hodges (1959), d'Ekman (1959), et de Lavallée et Hidiroglou (1988); de nouveau, son étude a montré que la stratification géométrique était la plus efficace parmi les méthodes comparées. Gunning, Horgan et Yancey (2004) ont appliqué cette méthode en vue de stratifier des populations comptables.

À l'instar de la méthode de la fonction cumulée de la racine carrée des fréquences, l'approche géométrique est une technique de stratification approximative, si bien que les

Mots clés : Stratification optimale; stratification géométrique; optimisation numérique; algorithme de Lavallée-Hidiroglou.

points de stratification qu'elle fournit peuvent s'écarter considérablement des points de stratification optimaux. Par ailleurs, il existe des approches, particulièrement pour la stratification unitaire, qui produisent des stratifications quasi optimales. Ces approches sont fondées sur l'utilisation d'algorithmes auto-appliqués ou de méthodes numériques d'optimisation pour produire les limites de strate (par exemple, Lavallée et Hidiroglou 1988; Lednicki et Wętczorkowski 2003; Kozak 2004). Toutefois, les méthodes de ce genre requièrent habituellement des limites initiales pour lancer le processus d'optimisation; les méthodes de stratification approximatives peuvent être utilisées pour rechercher ces points initiaux. Naturellement, les limites de strate initiales doivent être de haute qualité; sinon, l'optimisation risque de fournir un minimum local (Rivest 2002).

De nombreuses enquêtes comportent des variables d'intérêt positivement asymétriques. Le cas échéant, il est important de tenir compte de cet attribut lors de la stratification d'une population. Nombre de chercheurs ont essayé de créer des méthodes de stratification permettant de construire une strate dite « à tirage complet » (par exemple, Glasser 1962; Hidiroglou 1986) dont tous les éléments sont sélectionnés dans l'échantillon avec une probabilité égale à 1. Dans le contexte de l'échantillonnage stratifié, il s'agit du meilleur moyen de traiter les variables positivement asymétriques. Ces méthodes sont habituellement plus efficaces (de façon certaine, uniquement si une population est

Bibliographie

- Bickel, P.J., et Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.
- Booth, J.G., Butler, R.W. et Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282-1289.
- Folsom, R.E., Bayless, D.L. et Shah, B.V. (1971). Jackknifing for variance components in complex sample survey designs. *Proceedings of the Social Statistics Section, American Statistical Association*, 36-39.
- Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 181-184.
- Kovar, J.G., Rao, J.N.K. et Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16, Supplément, 25-45.
- McCarthy, P.J., et Snowden, C.B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics, Série 2*, 95, U.S. Government Printing Office.
- Public Health Service Publication, 85-1369, Washington, DC : U.S. Government Printing Office.
- Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.
- Royall, R.M., et Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.
- Royall, R.M., et Cumberland, W.G. (1981b). The finite population linear regression estimator: An empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- Rubin, D.B., et Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- Saigo, H., Shao, J. et Sitter, R.R. (2001). Bootstrap à demi-échantillon répété et répétitions équilibrées répétées en cas d'imputation aléatoire de données. *Techniques d'enquête*, 27, 209-218.
- Särndal, C.-E., Swenson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.
- Shao, J., et Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.
- Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. New York : Springer-Verlag.
- Sitter, R.R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.
- Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20, 135-154.

aucune expression analytique explicite de la variance pour les quantiles d'échantillon. En fait, aucune estimation de la variance n est associée aux estimations des quantiles de prix dans le rapport de l'ENP, tandis que les prix moyens sont publiés avec l'estimation de leur variance.

À la présente section, nous appliquons l'EBB abrégé aux données de l'ENP, en supposant que l'échantillonnage systématique peut être approximé par l'EASSR. Certaines strates ne contiennent qu'une seule UPE. En outre, $f_{1h} > n^{-1}$ dans certaines strates. Les strates de ce genre sont intégrées à des strates adjacentes de sorte que p_h , donnée par (3.1), soit comprise dans l'intervalle $[0, 1]$. Après regroupement, il existe plus de 280 strates. Nous supposons que l'effet de la reformulation des strates est négligeable.

Après reformulation des strates, nous employons l'EBB abrégé dans les strates composées de grandes villes. Par ailleurs, nous utilisons le bootstrap avec remise (Shao et Tu 1995, page 247) où la taille des répliques est $(n_h - 1)$ dans les strates composées de petites villes et de villages, où les fractions de sondage de premier degré sont faibles. Les estimations par quantile et leurs erreurs-types pour certains produits vendus par les petits points de vente sont présentées au tableau 4. Notons que les prix d'un produit donné sont discrets. Cependant, nous appliquons le bootstrap comme s'ils étaient continus. Cette approximation devrait être acceptable pour de nombreux produits, mais non pour ceux qui sont très bon marché, puisque, dans ce cas, un pourcentage élevé d'observations est concentré sur un prix particulier et l'erreur-type estimée peut être nulle.

données de l'ENP de 1997.

Les travaux du deuxième auteur ont été financés par la Japan Statistical Association. Ceux du troisième auteur ont été financés par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada. Les auteurs remercient le Bureau de la statistique, le ministère de la Gestion publique, des Affaires intérieures, des Postes et des Télécommunications, ainsi que le ministère de l'Economie, du Commerce et de l'Industrie du Japon d'avoir fourni les

Remerciements

au Japon.

Le bootstrap est utile pour estimer les variances dans le cas des enquêtes complexes, particulièrement lorsque plusieurs degrés ou les fractions de sondage sont grandes : l'EBB abrégé et l'EBB général. Dans les deux méthodes, une unité d'échantillonnage à un degré donné est soit retenue, soit remplacée avec une probabilité prédéterminée, afin de construire un échantillon bootstrap. L'EBB général a l'avantage de permettre le traitement de toute combinaison de tailles d'échantillon ≥ 2 , mais il nécessite plus de générations de nombres aléatoires que l'EBB abrégé. À titre d'illustration, nous avons appliqué l'EBB abrégé aux données de l'Enquête nationale sur les prix menée en 1997

6. Conclusion

Tableau 4
Quantités d'échantillon (erreurs-types) de certains produits pour les petits points de vente dans l'ENP

Produit	p	Quantité d'échantillon	(erreur-type)	Quantité d'échantillon	(erreur-type)	Quantité d'échantillon	(erreur-type)	Quantité d'échantillon	(erreur-type)	Quantité d'échantillon	(erreur-type)
Riz (5kg) ^a	0,10	239,4	(0,25)	255,2	(0,21)	278,3	(0,02)	299,1	(0,61)	914	(1,43)
Café instantané (1 flacon) ^b	(0,24)	714	(0,53)	788	(0,00)	536,8	(2,68)	549,4	(0,00)	346,5	(0,00)
Bière (24 cannettes) ^c	(0,13)	467,3	(0,40)	500,0	(0,82)	299,3	(0,00)	260,4	(0,00)	248,8	(0,00)
(10 yens)	(1,01)	(0,64)	(0,82)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
(100 yens)	(2,03)	(0,35)	(3,25)	(0,35)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Marques spécifiées : ^a Koshinikari; ^b Nescaré Gold Blend, 100g; ^c Sapporo (Nama) Black Label, 350ml.											

5. Application à l'Enquête nationale sur les prix menée en 1997 au Japon

L'objectif de l'ENP est d'analyser la formation des prix des principaux biens de consommation, comme les aliments, les vêtements et les appareils électroménagers. L'estimation par quantile joue un rôle essentiel dans cette analyse, et de nombreuses estimations par quantile fondées sur plusieurs stratifications a posteriori sont incluses dans les rapports de l'ENP.

L'échantillonnage stratifié à plusieurs degrés utilisé dans l'ENP de 1997 se résume comme suit :

Echantillonnage de premier degré. Ces UPE sont sélectionnées par BASSR indépendamment dans chaque strate. Le tableau 3 donne un aperçu des fractions de sondage de premier degré.

Echantillonnage de deuxième degré. Dans une municipalité sélectionnée, tous les grands points de vente sont dénombrés. Autrement dit, un échantillonnage en grappes à un degré est utilisé pour ces points de vente. Pour les petits points de vente, par contre, une municipalité échantillonnée est subdivisée en régions d'enquête (UEB), chacune constituée d'environ 100 points de vente. Un échantillonnage systématique est utilisé pour échantillonner les régions d'enquête. Les fractions de sondage au deuxième degré sont comprises entre 0,1 et 1,0.

Echantillonnage de troisième degré. Dans chaque région d'enquête sélectionnée, 40 points de vente (UFE) sont choisis par échantillonnage systématique ordonné en fonction du type de vente et du chiffre de ventes annuel déclaré lors du Recensement du commerce de 1994.

Tableau 3

Fractions de sondage de premier degré dans l'ENP de 1997

Catégorie de région	Taille de la population	N ^{ue} d'UPE	Fractions de l'échantillon	Table de sondage
Villes	≥ 100 000	221	2/3	1/1
Villes	50 000 – 99 999	220	1/3	2/3
Villes	< 50 000	224	1/5	1/3
Petites villes et villages	≥ 40 000	32	1/15	1/5
Petites villes et villages	< 40 000	2 536	1/15	1/15

À proprement parler, il n'existe aucune formule de variance valide pour les données de l'ENP, parce que celles-ci comportent un échantillonnage systématique. Cependant, pour estimer la variance, nous supposons que l'échantillonnage systématique peut être approximé par l'EASSR. Même sous cette condition simplifiée, il n'existe

Afin d'étudier les propriétés conditionnelles, nous avons ordonné les 1 000 exécutions de la simulation selon X/X_v et réparti les exécutions en 20 groupes de taille égale. Pour chaque groupe, nous avons calculé la moyenne de chaque estimateur de la variance. La figure 1 représente ces moyennes groupées pour chaque estimateur de la variance (sauf v_q puisqu'il présente un biais négatif important) en fonction de la moyenne groupée X/X_v , pour $p = 0,3$. L'EQM réelle est incluse dans le tracé également. Le graphique est semblable à celui utilisé par Royall et Cumberland (1981a, 1981b). Nous voyons que v_{EBB} suit l'EQM réelle, en grande partie comme v_{ewj} et v_{ma} , tandis que v_q ne le fait pas. Donc, l'EBB semble avoir une propriété conditionnelle désirable.

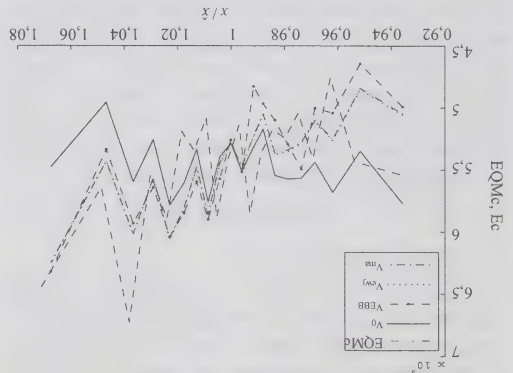


Figure 1. EQM et Ec(v) pour l'estimation par le ratio.

4.3 Estimation par quantile

Pour l'estimation par quantile, nous posons que $N = 100$, $n = 30$, $M_i = 100$ et $m_i = 10$, pour $i = 1, \dots, n$. Nous utilisons $B = 500$ répétitions bootstrap dans chacune des 5 = 5 000 exécutions de la simulation. Nous obtenons une approximation des EQM réelles au moyen de 50 000 exécutions de la simulation. Seuls les résultats pour v_{EBB} et v_{ewj} quand $p = 0,1$ sont résumés au tableau 2, parce que ceux obtenus quand $p = 0,3$ sont similaires. Nous voyons que la méthode de l'EBB donne d'assez bons résultats, avec un léger biais par excès, tandis que la méthode du jackknife pondérée extérieurement produit un biais important, à cause de son absence de convergence dans l'estimation de la variance pour les quantiles.

Tableau 2

Propriétés de v_{EBB} et v_{ewj} pour les quantiles 0,10, 0,25, 0,50, 0,75 et 0,90

v_{BBE}	v_{ewj}	Couverture (%)	Biais en CV	Couverture (%)
0,10	0,10	8,40	0,51	81,3
0,25	0,25	6,21	0,42	83,3
0,50	0,50	2,53	0,37	83,0
0,75	0,75	6,23	0,42	83,4
0,90	0,90	6,32	0,50	80,3

4. Une étude par simulation

À la présente section, nous décrivons l'exécution de simulations limitées pour étudier l'EBB dans le cas de l'estimation par le ratio et de l'estimation par quantile. Pour simplifier, nous considérons un EASSR à deux degrés et nous nous limitons à une seule strate.

4.1 Description générale de la simulation

Une population finie unistratifiée est générée selon la procédure qui suit et est maintenue fixe pour toutes les exécutions de la simulation afin d'observer les propriétés fondées sur le plan de sondage de l'EBB. Premièrement, la moyenne des variables auxiliaires dans la grappe i est générée par $\mu_i \sim N(\mu, \sigma^2)$ pour $i = 1, 2, \dots, N$. Puis, la variable auxiliaire x_{ik} de l'unité k dans la grappe i est générée par

(4.1)
$$x_{ik} = \mu_i + \varepsilon_{ik} \quad (k = 1, 2, \dots, M_i; i = 1, 2, \dots, N),$$

où $\varepsilon_{ik} \sim N(0, (1 - \rho)\sigma^2)$. La variable cible y_{ik} de l'unité k dans la grappe i est obtenue par

(4.2)
$$y_{ik} = a + bx_{ik} + e_{ik} \quad (k = 1, 2, \dots, M_i; i = 1, 2, \dots, N),$$

où $e_{ik} \sim N(0, \sigma^2/4)$. Les valeurs des paramètres sont fixées à $\mu = 100, \sigma = 10, \rho = 0, (0, 3), a = 0$ et $b = 1$, et l'EASSR à deux degrés est utilisé tout au long de l'étude par simulation.

4.2 Estimation par le ratio

Soit $N = 50, n = 15, M_i = 20$ et $m_i = 3$, for $i = 1, \dots, n$. Considérons l'estimateur par le ratio du total de population, Y_i ,

$$\hat{Y}_R = R X_i,$$

où $X_i = \sum_{h=1}^{N_i} \sum_{k=1}^{M_{hi}} x_{ik}$ est le total de population des x , $R = Y/X, \hat{Y}_R = \sum_{h=1}^{N_i} \sum_{k=1}^{M_{hi}} \hat{Y}_{hk}, \hat{X} = \sum_{h=1}^{N_i} \sum_{k=1}^{M_{hi}} \hat{X}_{hk} = \sum_{h=1}^{N_i} (N_h/n^h) \sum_{k=1}^{M_{hi}} \hat{X}_{hk}$ et $X_{hi} = (M_{hi}/m_{hi}) \sum_{k=1}^{m_{hi}} X_{hik}$.

Aux fins de comparaison, nous considérons un certain nombre d'estimateurs de la variance utilisables dans ce simple contexte :

1) L'estimateur classique de la variance est dénoté

$$v_0(\hat{Y}_R) = N^2 \frac{1 - f_1}{n} \frac{\sum_i (Y_i - R \hat{X}_i)^2}{n - 1} + \frac{n}{N} \sum_i \frac{m_i}{M_i^2 (1 - f_{2i}) (S_{2i}^{D_{2i}})^2},$$

où $f_1 = n/N, f_{2i} = m_i/M_i$ et $S_{2i}^{D_{2i}} = \sum_j (Y_j - R X_j)^2 / (m_i - 1)$.

2) L'estimateur par le jackknife avec suppression d'une

UPB à la fois corrigé pour la fraction de sondage de premier degré, parfois utilisé même s'il n'est pas entièrement correct, est dénoté

Tableau 1 Comparaison des estimateurs de la variance pour \hat{Y}_R				
p	Biais en %	CV	Couverture (90 %)	
0.1	v_0 -1,70	0,28	89,2	v_{EBB} -0,62
	v_{EWJ} -0,33	0,30	88,9	v_{EWJ} -0,33
	v_{EJ} -26,55	0,39	80,5	v_{EJ} -26,55
	v_{ma} -0,39	0,30	89,4	v_{ma} -0,39
0.3	v_0 -0,67	0,28	86,6	v_{EBB} -1,63
	v_{EWJ} -0,74	0,29	86,5	v_{EWJ} -0,74
	v_{EJ} -26,85	0,39	80,2	v_{EJ} -26,85
	v_{ma} -0,87	0,29	86,4	v_{ma} -0,87

Nous voyons au tableau 1 que v_{EBB}, v_0, v_{EWJ} et v_{ma} donnent des résultats comparables et bons, excepté que le coefficient de variation (cv) des méthodes de rééchantillonnage est un peu plus élevé que celui des méthodes sans rééchantillonnage, ce qui est typique. Le jackknife avec suppression d'une UPB à la fois donne des résultats médiocres.

des intervalles de confiance à 90 %, comme mesures de leur variance, ainsi que les probabilités empiriques de couverture, coefficient de variation des divers estimateurs de la Monte Carlo du biais relatif en pourcentage et du exécutions de la simulation et nous utilisons les estimations des EQM réelles d'après 10 000 une approximation des EQM réelles d'après 10 000 des $S = 1\,000$ exécutions de la simulation. Nous obtenons $B = 100$ répriques bootstrap dans chacune Nous utilisons $B = 100$ répriques bootstrap dans chacune

équation (8.10.6)) de la forme

(4.6)
$$v_{EWJ}(\hat{Y}_R) = (X/Y)^2 v_0(\hat{Y}_R).$$

être dérivé sous la forme

(4.4)
$$v_{EJ}(\hat{Y}_R) = (1 - f_1) \frac{n}{n - 1} \sum_i (Y_i^{R(i)} - \bar{Y}^{R(i)})^2,$$

où $\bar{Y}^{R(i)}$ est l'estimateur recalculé après l'élimination de la i^{e} UPB et $\bar{Y}^{R(i)} = \sum_i Y^{R(i)} / n$.

3) Un estimateur par le jackknife pondéré extérieurement (voir Folsom, Bayless et Shah 1971) qui comprend une correction pour les deux degrés d'échantillonnage peut

Dénoter l'ensemble candidat par $\{UPE_{hi} : i = 1, 2, \dots, n_h - 1\}$. Pour chaque UPE i dans l'échantillon de la strate h : a) la garder dans l'échantillon bootstrap avec la probabilité

$$p_h = 1 - \frac{1}{1 - f_{1h}} \frac{2}{2(1 - n_h^{-1})}; \quad (3.4)$$

ou b) la remplacer par une autre sélectionnée au hasard à partir de $\{UPE_{hi} : i = 1, 2, \dots, n_h - 1\}$. Si l'option est a), passer à l'étape II'.

Étape II'.

Pour l'unité hi retenue à l'étape I', tirer $(m_{hi} - 1)$ UPE par EAS avec remise parmi les m_{hi} UPE dans l'UPE hi . Dénoter l'ensemble candidat par $\{USE_{hi} : j = 1, 2, \dots, m_{hi} - 1\}$. Pour chaque USE hi dans l'UPE hi retenue à l'étape I' : c) la garder dans l'échantillon bootstrap avec la probabilité

$$q_{hi} = 1 - \frac{1}{1 - f_{1h}} \frac{2}{2p_h} \frac{1}{(1 - f_{2hi})}; \quad (3.5)$$

ou d) la remplacer par une autre sélectionnée au hasard à partir de $\{USE_{hi} : j = 1, 2, \dots, m_{hi} - 1\}$. Si l'option est c), passer à l'étape III'.

Étape III'.

Pour l'unité hi retenue à l'étape II', tirer $l_{hi} - 1$ UPE par EAS avec remise parmi les l_{hi} UPE dans l'USE hi dans l'UPE hi . Dénoter l'ensemble candidat par $\{UFE_{hi} : k = 1, 2, \dots, l_{hi} - 1\}$. Pour chaque UFE hi dans l'USE hi dans l'UPE hi : e) la garder dans l'échantillon bootstrap avec la probabilité

$$r_{hi} = 1 - \frac{1}{1 - f_{1h}} \frac{2}{2p_h} \frac{1}{f_{2hi}} \frac{q_{hi}}{(1 - l_{hi}^{-1})}; \quad (3.6)$$

ou f) la remplacer par une autre sélectionnée au hasard à partir de $\{UFE_{hi} : k = 1, 2, \dots, l_{hi} - 1\}$.

Il est facile de voir que $p_h, q_{hi}, r_{hi} \in [0, 1] \forall n_h, m_{hi}, l_{hi} \geq 2$.

La raison justifiant la sélection aléatoire d'un ensemble candidat dans l'EBB général est la suivante. Pour fixer les idées, considérons un EASSR à un degré dans une seule strate. Soit \bar{y} une moyenne d'échantillon bootstrap sous l'EBB abrégé avec une probabilité arbitraire $p \in [0, 1]$. Alors, nous pouvons montrer que $V^*(\bar{y}) = n^{-1}(1 - n^{-1})s^2(1 - p^2)$, où $s^2 = \sum_i (y_i - \bar{y})^2 / (n - 1)$. Notons que $V^*(\bar{y})$ est monotone décroissante par rapport à p dans l'intervalle $[0, 1]$. Donc, $\min_{p \in [0, 1]} V^*(\bar{y}) = 0$ et $\max_{p \in [0, 1]} V^*(\bar{y}) = n^{-1}(1 - n^{-1})s^2$. Si $f_1^p > n^{-1}$, puis $\max_p V^*(\bar{y}) < v(\bar{y})$. La notion clé de l'EBB général est que nous pouvons rendre $\max_p V^*(\bar{y})$ plus grand que $v(\bar{y})$ en introduisant une

Il n'est pas difficile d'étendre l'approche de l'EBB à des plans comportant plus de trois degrés. Par exemple, pour un plan stratifié à quatre degrés, une UFE au quatrième degré dans la strate h est retenue avec la probabilité

$$\sqrt{1 - p_h} \frac{1}{1 - f_{1h}} \frac{q_{hi}}{f_{2hi}} \frac{1}{f_{3hi}} \frac{1}{(1 - g_{hjk})} \frac{1}{(1 - f_{4hjk})}$$

ou remplacée dans l'EBB abrégé, où g_{hjk} est la fraction de l'échantillon de quatrième degré et f_{4hjk} est la fraction de sondage de quatrième degré. Les extensions ultérieures sont analogues.

L'EBB général randomise un ensemble candidat simplement pour remédier à l'infaisabilité de l'EBB abrégé. Ce concept présente des similarités avec le bootstrap approximativement bayésien de Rubin et Schenker (1986).

Un inconvénient de l'EBB général comparativement à l'EBB abrégé est que le premier nécessaire, en moyenne, $\sum_h \{(n_h - 1) + p_h \sum_i (m_{hi} - 1) + p_h \sum_i q_{hi} \sum_j (l_{hij} - 1)\}$ générations de nombres aléatoires de plus que le second, où p_h, q_{hi} , et r_{hi} sont donnés par (3.4), (3.5) et (3.6), lorsque les tailles d'échantillon et (ou) le nombre de strates sont grands. Afin de réduire les générations de nombres aléatoires dans l'EBB général, on peut créer un ensemble candidat en supprimant aléatoirement une unité de l'échantillon original et utiliser

$$p_h = (m_h + 1/2 - \sqrt{(m_h + 1/2)^2 - n_h(1 + f_{1h})}) \quad (3.7)$$

$$q_{hi} = (m_{hi} + 1/2 - \sqrt{(m_{hi} + 1/2)^2 - f_{1h} p_h^{-1} m_{hi}(1 + f_{2hi})}) \quad (3.8)$$

$$r_{hi} = (l_{hij} + 1/2 - \sqrt{(l_{hij} + 1/2)^2 - f_{1h} p_h^{-1} f_{2hi} q_{hi}^{-1} l_{hij}(1 + f_{3hi})}) \quad (3.9)$$

au lieu des trois équations susmentionnées. On peut montrer que $p_h, q_{hi}, r_{hi} \in [0, 1]$. La preuve de cette version modifiée de l'EBB général est similaire.

l'échantillon d'UPE est $n = \sum_{i=1}^m n_i$. Au deuxième degré, un échantillon de m_{hi} unités secondaires d'échantillonnage (USE) est sélectionné à partir de l'UPE i de taille M_{hi} dans la strate h par EASSR. Au troisième degré, un échantillon de l_{hi} unités finales d'échantillonnage (UFE) est sélectionné à partir de l'USE ij de taille L_{hi} dans la strate h par EASSR. Un vecteur de mesures de certaines caractéristiques des unités est représenté par $\mathbf{y}^{hijk} = (y^{hijk}_1, y^{hijk}_2, \dots, y^{hijk}_k)^T$, où les indices inférieurs $hijk$ sont l'étiquette de strate, l'étiquette d'UPE, l'étiquette d'USE et l'étiquette d'UFE, respectivement. Le paramètre de population d'intérêt $\theta = \theta(S)$, où $S = \{y^{hijk} : h = 1, 2, \dots, H; i = 1, 2, \dots, N_h; j = 1, \dots, M_{hi}; k = 1, \dots, L_{hi}\}$, est habituellement estimé par $\hat{\theta} = \theta(s)$, où $s = \{y^{hijk} : h = 1, 2, \dots, H; i = 1, 2, \dots, n_h; j = 1, \dots, m_{hi}; k = 1, \dots, l_{hi}\}$. Le vecteur des totaux de population est dénoté $\mathbf{Y} = (Y_1, \dots, Y_T)^T$. Ici, son estimateur sans biais est :

$${}^{\epsilon}!q \sum_{\wedge}^{I=7} ({}^y u / {}^y N) \sum_H^{I=4} = {}^q \sum_{\wedge}^{I=4} = \sum_{\wedge}$$

où $\mathbf{X}_i \mathcal{M}_i = \sum_{j=1}^K w_{ij} \mathbf{X}_j \mathcal{M}_j$ et $\mathbf{X}_i \mathbf{M}_i = (L)^{\mathbf{X}_i / (1/n_i)}$ peut s'écrire sous la forme $\mathbf{X}_i \mathbf{M}_i = (L)^{\mathbf{X}_i / (1/n_i)} \mathcal{M}_i (L)^{1/n_i}$, où $w_{ij} \mathbf{X}_j \mathcal{M}_j = \mathbf{X}_j \mathcal{M}_j$, $\mathbf{X}_j \mathcal{M}_j = \mathbf{X}_j \mathbf{M}_j$, $\mathbf{X}_j \mathbf{M}_j = (L)^{\mathbf{X}_j / (1/n_j)}$ et $\mathcal{M}_j = (L)^{1/n_j}$, une estimation sans biais de $\text{Var}(Y)$ est

$$\frac{1}{f_2^2} \sum_{u=1}^f \frac{u}{M} \sum_{N=1}^f \frac{u}{N} + \frac{1}{f_2^2} \sum_{N=1}^f \frac{u}{N} = \left(\frac{f}{f_2} \right)^2$$

avec

$$\begin{aligned} \tilde{f}_h^k &= n_h^{-1} \sum_{i=1}^n \tilde{f}_h^k(\tilde{X}_{hi}^k) = m_h^{-1} \sum_{i=1}^n \tilde{f}_h^k(\tilde{X}_{hi}^k) / L_h^{hjk}, \quad \tilde{f}_h^{jk} = l_h^{hjk} \sum_{i=1}^n \tilde{f}_h^k(\tilde{X}_{hi}^k) / L_h^{hjk}, \quad \tilde{f}_h^k = l_h^{hk} \sum_{i=1}^n \tilde{f}_h^k(\tilde{X}_{hi}^k) / L_h^{hk}, \\ \tilde{f}_h^k &= n_h^{-1} / N_h^{jk}, \quad \tilde{f}_h^{jk} = m_h^{-1} / M_h^{jk}, \quad \tilde{f}_h^k = l_h^{hk} / L_h^{hk}, \quad \tilde{f}_h^{jk} = l_h^{hjk} / L_h^{hjk}, \quad \tilde{f}_h^k = \sum_{i=1}^n (\tilde{X}_{hi}^k - \bar{\tilde{X}}_h^k)^2 / s_h^2, \\ \tilde{f}_h^k &= \sum_{i=1}^n (\tilde{X}_{hi}^k - \bar{\tilde{X}}_h^k)^2 / (m_h - 1), \quad \tilde{f}_h^{jk} = \sum_{i=1}^n (\tilde{X}_{hi}^k - \bar{\tilde{X}}_h^k)^2 / (m_h - 1), \quad \tilde{f}_h^k = \sum_{i=1}^n (\tilde{X}_{hi}^k - \bar{\tilde{X}}_h^k)^2 / s_h^2, \end{aligned}$$

(Särndal, Swensson et Wretman 1992, pages 148-149).

1992, pages 148-149).

3. Bootstrap de type Bernoulli proposé

Afin de traiter la question de l'échantillonnage à plusieurs degrés dans une strate, nous nous proposons un bootstrap à plusieurs degrés. Pour simplifier les idées, nous commençons par introduire une version simple, dont l'application présente certaines limites. Puis, nous décrivons une forme plus générale qui permet d'éviter ces difficultés.

EBB abrégé

Etape 1. Pour chaque UPE de l'échantillon, h_i , dans la strate $h, h = 1, \dots, H$: a) la garder dans l'échantillon bootstrap avec la probabilité

Funaoka, Saigo, Sitter et Toida : Bootstrap de type Bernoulli pour l'échantillonnage stratifié à plusieurs degrés

$$(3.1) \quad \frac{(1^y u - 1)}{(1^{y_1} u - 1)} - 1 = d^y$$

ou b) la remplacer par une autre sélectionnée au hasard parmi les n_h UPE. Si l'option est a), passer à l'étape II.

Etape II.

strate h retenue à l'étape I : c) la garder dans l'échantillon bootstrap avec la probabilité

$$(3.2) \quad \left(\frac{(1 - f_{2h}^y)}{(1 - f_{1h}^y)} \right)^d = 1 - q_{h1}^y$$

ou d) la remplacer par une autre sélectionnée au hasard parmi les m_{hi} USE dans l'UPE hi de la strate h . Si l'option est c), passer à l'étape III.

Etape III.

Pour chaque UFE h_{ij} dans l'USE h_{ij} dans l'UPE h_i de la strate h : e) la garder dans l'échantillon bootstrap avec la probabilité

$$(3.3) \quad r_{ij} = \sqrt{1 - \frac{f_{1h} f_{2hi} f_{3hj} b_{-1} d_{-1}}{(1-f_{3hj})(1-l_{-1})(1-l_{ij})}};$$

(ou f) la remplacer par une autre sélectionnée au hasard parmi les l_{hy} UFE dans l'USE hy dans l'UPE hi de la strate h .

Soit K_{hj}^* le nombre de fois que l'unité $hijk$ figure dans la réplique bootstrap, alors, l'estimation du total par le bootstrap est $\hat{Y}^* = \sum_{hjk} w_{hj}^* Y_{hijk}$, où $w_{hj}^* = K_{hj}^* w_{hj}^*$, et l'estimation de $V(\hat{\theta})$ par le bootstrap est $\hat{v}^*(\hat{\theta}) = V^*(\hat{\theta})$, où $\hat{\theta}^* = \theta(\hat{Y}^*)$ et V^* représente la variance sous la procédure de rééchantillonnage. Habituellement, l'estimation de la variance par le bootstrap est obtenue par simulation de Monte Carlo. Autrement dit, on répète les étapes I à III un grand nombre de fois, B , pour obtenir $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ et on utilise

$$\sum_{B=1}^q (\hat{\theta}_*^{(q)} - \hat{\theta}_*^{(q)})^2 / B = (\hat{\theta}_*^{(q)})^2$$

où $\theta_i = \sum_{j=1}^p \theta_j / B$. Dans la plupart des cas, il est possible de remplacer θ_i par θ . Cela permet au méthodologiste d'enquêter de créer un ensemble de poids de rééchantillonnage w_{HT}^i pour chaque bootstrap et de les inclure dans les fichiers de données à grande diffusion.

Il est clair que l'FBB abrégé n'est applicable que si $q_{h_i}^i, r_{h_i}^i \in [0, 1] \forall h_i, i$. Par exemple, il est nécessaire que $f_{h_i}^i \geq n_i$. Nous pouvons modifier chaque étape et changer $p_{h_i}^i, q_{h_i}^i, r_{h_i}^i$ en conséquence.

EBB général

Étape I. Tirer $(n_h - 1)$ UPE par EAS avec remise parmi les n_h UPE de l'échantillon, $h = 1, \dots, H$.

Bootstrap de type Bernoulli pour l'échantillonnage stratifié à plusieurs degrés

Fumio Funakoka, Hiroshi Saigo, Randy R. Sitter et Tsutomu Toida¹

Résumé

Nous proposons dans cet article une méthode de bootstrap de type Bernoulli facilement applicable à des plans stratifiés à plusieurs degrés où les fractions de sondage sont grandes, à condition qu'un échantillonnage aléatoire simple sans remise soit utilisé à chaque degré. La méthode fournit un ensemble de poids de rééchantillonnage qui donnent des estimations convergentes de la variance pour les estimateurs lisses ainsi que non lisses. La force de la méthode tient à sa simplicité. Elle peut être étendue facilement à n'importe quel nombre de degrés d'échantillonnage sans trop de complications. L'idée principale est de garder ou de remplacer une unité d'échantillonnage à chaque degré d'échantillonnage en utilisant des probabilités prédéterminées pour construire l'échantillon bootstrap. Nous présentons une étude par simulation limitée afin d'évaluer les propriétés de la méthode et, à titre d'illustration, nous appliquons cette dernière à l'Enquête nationale sur les prix menée en 1997 au Japon.

Mots clés : Enquête complexe; linéarisation; quantiles; rééchantillonnage; stratification.

1. Introduction

De nombreuses enquêtes à grande échelle sont réalisées selon un plan d'échantillonnage stratifié à plusieurs degrés. Or, lorsqu'on utilise ce genre de plan, l'estimation de la variance peut être analytiquement complexe, voire même impossible. En outre, pour les ensembles de données à grande diffusion, les formes particulières des estimateurs dont l'utilisateur pourrait souhaiter se servir pour obtenir les estimations de la variance sont inconnues. Par conséquent, des méthodes de rééchantillonnage sont souvent utilisées pour produire un ensemble de poids de rééchantillonnage qui peuvent être fournis avec l'ensemble de données et utilisés en vue d'estimer la variance pour une grande gamme d'estimateurs possibles. Le bootstrap est particulièrement utile, puisqu'il permet de traiter des statistiques d'échantillon lisses ainsi que non lisses sous des plans d'échantillonnage à plusieurs degrés. Un sommaire de plusieurs méthodes du bootstrap pour l'échantillonnage en population finie peut être consulté dans Shao et Tu (1995, pages 232-282) (voir aussi, Gross 1980; Bickel et Freedman 1984; McCarthy et Snowden 1985; Rao et Wu 1988; Kovar, Rao et Wu 1992a, b; Booth, Buiter et Hall 1994; Shao et Sitter 1996).

Si la fraction de sondage de premier degré est faible, diverses méthodes du bootstrap existent pour traiter l'échantillonnage de premier degré comme s'il avait eu lieu avec remise afin d'estimer la variance. Dans le cas où les fractions de sondage de premier degré ne sont pas négligeables, un moins grand nombre de résultats sont disponibles. Pour le « bootstrap » sous échantillonnage à deux degrés avec échantillonnage aléatoire simple (EAS) à

chaque degré, voir Sitter (1992a, 1992b) et pour celui avec probabilités inégales, voir Rao et Wu (1988). Cependant, si les fractions de sondage de premier degré ne sont pas négligeables, aucune méthode du bootstrap simple n'existe pour trois degrés ou plus d'échantillonnage. Dans le présent article, nous proposons une nouvelle méthode du bootstrap qui permet de traiter facilement les cas pour lesquels l'échantillonnage aléatoire simple (EAS) est utilisé à chaque degré. Nous l'appelons bootstrap de type Bernoulli (EBB) à cause de sa ressemblance à l'échantillonnage à partir d'une loi de Bernoulli. Nous utilisons les données de l'Enquête nationale sur les prix (ENP) du Japon pour l'illustrer. Le plan de l'article est le suivant. À la section 2, nous présentons la notation pour l'échantillonnage stratifié à trois degrés. À la section 3, nous décrivons deux types d'EBB. À la section 4, nous étudions les propriétés de la méthode par simulation. À la section 5, nous décrivons le plan d'échantillonnage de l'ENP de 1997 et illustrons l'application de l'EBB aux données de l'ENP. Enfin, à la section 6, nous présentons nos conclusions.

2. Échantillonnage stratifié à trois degrés

Dans l'échantillonnage aléatoire stratifié, la population finie, constituée de N unités primaires d'échantillonnage (UP), est fractionnée en H strates non chevauchantes contenant N_1, N_2, \dots, N_H UP, respectivement, donc, $\sum_{h=1}^H N_h = N$. Un échantillon aléatoire simple sans remise (EASSR) d'UP est tiré indépendamment dans chaque strate. Les tailles d'échantillon dans chaque strate sont dénotées par n_1, n_2, \dots, n_H , et la taille totale de

1. F. Funakoka, professeur, Faculty of Economics, Shinshu University, 3-1-1 Asahi, Matsumoto, Nagano, 390-8621, Japon; H. Saigo, professeur, School of Political Science and Economics, Waseda University, 1-6-1 Nishitwaseda Shingaku, Tokyo, 169-8050, Japon; R.R. Sitter, professeur, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, B.C., V5A 1S6, Canada; T. Toida, professeur associé, Faculty of Social and Information Studies, Gunma University, 2-4 Aramakicho, Maebashi, Gunma 371-8510, Japon.

Remerciements

La présente étude a été financée par la bourse ITR-0427889 de la National Science Foundation. Les auteurs remercient le rédacteur adjoint et les examinateurs de leurs commentaires et suggestions.

Bibliographie

Barnard, J., et Meng, X. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8, 17-36.

Barnard, J., et Rubin, D.B. (1999). Small-sample degrees of freedom with multiple-imputation. *Biometrika*, 86, 948-955.

Collins, L.M., Schafer, J.L. et Kam, C.K. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, 6, 330-351.

Gelfand, A.E., et Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Heitjan, D.F., et Little, R.J.A. (1991). Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics*, 40, 13-29.

Kernickell, A.B. (1998). Multiple imputation in survey of consumer finances. Dans *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 11-20.

Li, K.H., Raghunathan, T.E. et Rubin, D.B. (1991a). Large-sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86, 1065-1073.

Li, K.H., R.T.E., Meng, X.T. et Rubin, D.B. (1991b). Significant levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, 1, 65-92.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Little, R.J.A., et Raghunathan, T.E. (1997). Should imputation of missing data condition on all observed variables? Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 617-622.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9, 8.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science*, 9, 538-558.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. et Solenberger, P. (2001). Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression. *Techniques d'enquête*, 27, 91-103.

Raghunathan, T.E., and Paulin, G.S. (1998). Multiple imputation of income in the Consumer Expenditure Survey: Evaluation of statistical inference. Dans *Proceedings of the Section on Business and Economic Statistics*, American Statistical Association, 1-10.

Raghunathan, T.E., Reiter, J.P. et Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T.E., et Siscovick, D.S. (1996). A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45, 335-352.

Reiter, J.P. (2003). Inférence pour les ensembles de microdonnées à grande diffusion partiellement synthétiques. *Techniques d'enquête*, 29, 203-211.

Reiter, J.P. (2004). Utilisation simultanée de l'imputation multiple pour les données manquantes et le contrôle de la divulgation. *Techniques d'enquête*, 30, 263-271.

Reiter, J.P. (2005). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 168, 185-205.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York : John Wiley & Sons, Inc.

Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462-468.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A. et Rubin, D.B. (1998). The NHANES III multiple imputation project. Dans *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 28-37.

Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G. et Cohen, A.J. Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association*, a paraître.

omises. Les imputeurs qui soupçonnent l'existence de relations de ce genre devraient inclure les interactions appropriées avec les variables nominales pour les caractéristiques du plan, comme nous l'avons fait dans l'exemple des enquêtes NHANES. Dans le cas de certaines enquêtes, le plan peut être si complexe qu'il est impossible d'inclure des variables nominales pour chaque grappe. Le cas échéant, les imputeurs peuvent simplifier le modèle en ce qui concerne les variables du plan, par exemple en regroupant des catégories de grappes ou en incluant des variables de substitution (par exemple, taille de grappe) qui sont corrélées à la variable d'enquête d'intérêt.

Les simulations donnent à penser qu'il pourrait être avantageux d'utiliser des modèles hiérarchiques plutôt que des modèles à effets fixes pour l'imputation des données manquantes, particulièrement lorsque les effets de grappe sont semblables. Toutefois, les modèles hiérarchiques sont plus difficiles à ajuster que les modèles à effets fixes. Ainsi, l'ajustement de modèles hiérarchiques dans le cas de plans d'échantillonnage complexes lorsque des données manquent pour plusieurs variables continues et catégoriques est une tâche redoutable. Des modèles hiérarchiques séquentiels pourraient peut-être être ajustés dans un esprit semblable aux imputations par régression séquentielle de Raghunathan et coll. (2001). Il s'agit d'un domaine dans lequel devraient se poursuivre les travaux de recherche. Un autre inconvénient des modèles hiérarchiques est qu'il est plus facile de les spécifier incorrectement que les modèles à effets fixes. Ainsi, si les effets de grappe suivent une loi non normale, le modèle hiérarchique normal utilisé dans le présent article pourrait donner des imputations non plausibles.

Dans le cas de l'imputation multiple, la clé du succès réside dans la spécification d'un modèle d'imputation qui décrit raisonnablement la loi conditionnelle des valeurs manquantes sachant les valeurs observées. Souvent, les caractéristiques du plan sont corrélées aux variables d'enquête, de sorte que leur inclusion dans les modèles d'imputation réduit les risques d'erreur de spécification du modèle. Nous pensons que, dans de nombreux cas, les biais que peut causer l'exclusion de variables importantes, du plan ou d'autres variables reliées au mécanisme de création des données manquantes, surpassent les inefficacités qui pourraient résulter de l'estimation de petits coefficients. Cela renforce le conseil général fréquemment donné concernant l'imputation multiple : inclure toutes les variables qui sont reliées aux données manquantes dans les modèles d'imputation afin de rendre ignorable le mécanisme de création des données manquantes (par exemple, Meng 1994; Little et Raghunathan 1997; Schafer 1997; Collins, Schafer et Kam 2001).

Le tableau 5 donne les résultats pour les deux stratégies d'imputation. Pour toutes les analyses, les deux ensembles d'estimations sont fort semblables. Dans ce cas, l'intégration des variables du plan dans le modèle d'imputation n'a presque pas d'effets sur les résultats. Cela tient, en partie, aux faibles fractions d'information manquante et à l'insignifiance relative des effets de strate et de grappe. Cependant, la pénalité pour l'inclusion des caractéristiques du plan dans le modèle d'imputation est minime. À la lumière des résultats des simulations présentés à la section 3, nous intégrerions les caractéristiques du plan dans ce modèle d'imputation.

Tableau 5
Comparaison des résultats des données réelles lorsque les caractéristiques du plan sont incluses dans le modèle d'imputation et lorsque les caractéristiques du plan sont ignorées

Moyenne de BP060		Est. ponc.		IC à 95 %	
Variables du plan	0,319	E-t.			
Pas de variable du plan	0,319	(0,299, 0,339)			
Ordonnée à l'origine : régression logarithue					
Variables du plan	0,362	(0,256, 0,467)			
Pas de variable du plan	0,352	(0,251, 0,454)			
Pente : régression logarithue					
Variables du plan	-0,409	(-0,449, -0,369)			
Pas de variable du plan	-0,407	(-0,444, -0,371)			

5. Conclusion

Quoique limitées, les études par simulation donnent à penser que ne pas tenir compte du plan d'échantillonnage dans l'imputation multiple peut être une pratique risquée. Lorsque les variables du plan sont corrélées aux variables d'enquête, comme dans notre simulation A, omettre de les inclure peut donner lieu à un biais important. Par ailleurs, l'inclusion de variables du plan non pertinentes, comme dans notre simulation B et dans l'exemple des enquêtes NHANES, produit, au pire, des inférences inefficaces et prudenées, lorsque les modèles d'imputation sont par ailleurs spécifiés correctement.

Inclure des variables nominales pour les effets de grappe réduit considérablement le biais comparativement à la non-prise en compte totale du plan. Cependant, l'introduction aveugle de variables nominales n'est pas une solution automatique. Lorsque la pente de la régression ou les variances diffèrent selon la grappe, l'utilisation de la méthode FX ou HM peut produire des estimations biaisées, puisque des caractéristiques importantes du plan sont

Reiter, Raghunathan et Kinney : L'importance de la modélisation du plan d'échantillonnage dans l'imputation

Tableau 4

La moyenne de population est égale à 0,34 et les coefficients de régression de population sont égaux à 0,14 et 10,04

Méthode	Couv. IC à 95 %	Est. ponc.	Var.	Var. est.	Var. (var. est.)	EQM (var. est.)
Moyenne de Y	Données complètes	94,7	0,35	14,61	14,73	32,65
	EAS	95,7	0,12	16,45	19,22	40,65
	FX	97,8	0,40	19,64	28,29	97,66
	HM	95,1	0,26	18,77	19,16	47,29
	Données complètes	93,7	0,12	7,13	7,20	5,31
	EAS	96,8	-0,10	8,97	11,72	13,59
	FX	98,6	0,17	12,23	20,62	39,84
	HM	96,2	0,03	10,45	11,61	15,09
	Données complètes	94,5	10,04	0,07	0,07	0,001
	EAS	96,4	10,07	0,10	0,13	0,002
	FX	96,4	10,04	0,12	0,15	0,003
	HM	95,2	10,05	0,11	0,12	0,002
Ordonnée à l'origine	Données complètes	94,7	0,35	14,61	14,73	32,65
	EAS	95,7	0,12	16,45	19,22	40,65
	FX	97,8	0,40	19,64	28,29	97,66
	HM	95,1	0,26	18,77	19,16	47,29
	Données complètes	93,7	0,12	7,13	7,20	5,31
	EAS	96,8	-0,10	8,97	11,72	13,59
Pente	Données complètes	94,5	10,04	0,07	0,07	0,001
	EAS	96,4	10,07	0,10	0,13	0,002
	FX	96,4	10,04	0,12	0,15	0,003
	HM	95,2	10,05	0,11	0,12	0,002

La méthode EAS produit enfin des estimations ponctuelles dont les moyennes sont comprises dans la marge d'erreur de simulation de l'estimation ponctuelle moyenne d'après des données complètes. Il en est ainsi parce que l'imputation sous EAS reflète raisonnablement bien la structure de population. Il semble donc que ne pas tenir compte du plan d'échantillonnage dans les modèles d'imputation peut fournir des inférences acceptables lorsque

Pour la méthode FX, le pourcentage d'intervalles de confiance qui couvrent \hat{Q} est plus grand que le pourcentage observé pour les intervalles calculés d'après des données complètes et pour la méthode HM. Cela tient au fait que la variance estimée pour FX a tendance à être plus grande que la variance réelle. Ce biais par excès apparaît dans T_M existe également dans le cas de la méthode EAS, ce qui donne un pourcentage de couverture plus grand que celui calculé pour les données complètes et la méthode HM.

4. Exemple fondé sur des données réelles

Nous allons maintenant examiner l'effet de la prise en compte de la stratification et de la mise en grappes lors de

L'imputation pour traiter les données manquantes dans un fichier à grande diffusion des Nations Health and Nutrition Examination Surveys réalisées de 1999 à 2002. Les individus sont groupés en 56 grappes réparties entre 82 strates. De 5 % à 10 % de données manquantes sont relevées pour de nombreuses variables.

premier est le pourcentage des personnes dans la population qui ont déjà fait vérifier leur taux de cholestérol (BPQ060). La proportion de données manquantes pour cette variable est d'environ 15 %. Les deuxième et troisième sont les coefficients de régression de population dans une régression logistique de BPQ060 sur le ratio revenu-seuil de pauvreté de la famille (INDFMPIR), variable continue pour laquelle la proportion de valeurs manquantes est d'environ 12 %. Ces paramètres sont estimés par des méthodes fondées sur

Les imputations fondées sur la méthode EAS produisent des estimations gravement biaisées et une couverture très médiocre des intervalles de confiance dans cette population. Ces problèmes existent même si peu d'information manque et malgré le fait que nous utilisons des estimateurs sans biais par rapport au plan de sondage pour les inférences. Les méthodes FX et HM produisent toutes deux des estimations ponctuelles qui concordent approximativement avec les estimations ponctuelles basées sur les données complètes et donnent toutes deux des taux de couverture qui correspondent approximativement aux taux obtenus pour l'inférence d'après les données complètes. FX et HM ont des profils similaires, parce que les modèles à effets fixes et les modèles hiérarchiques produisent des estimations similaires des paramètres dans l'équation 5.

Lors de l'estimation de la moyenne de population, la variance associée à FX ou à HM n'est que légèrement plus grande que celle associée à l'estimateur d'après des données complètes. Il en est ainsi à cause des grands effets de grappe, qui font de la variance dans les cellules d'imputation un facteur dominant relativement à la variance entre cellules d'imputation. Autrement dit, la fraction d'information manquante due aux données manquantes est relativement faible comparativement à l'effet de la mise en grappes.

3.2 Simulation B : Illustration de l'inclusion de variables explicatives non pertinentes

La modélisation des caractéristiques du plan est essentielle quand ces dernières sont reliées aux variables d'enquête d'intérêt. Quelle est l'incidence de la modélisation de caractéristiques non pertinentes du plan sur les inférences? À la présente section, nous présentons les résultats de deux études par simulation réalisées en vue d'étudier cette question.

Tableau 3
Propriétés des procédures d'imputation lorsque la population comprend des effets de strate, mais non des effets de grappe
La moyenne de population est égale à 0,34 et les coefficients de régression de la population sont égaux à 0,14 et 10,13

Méthode	Couv. IC à 95 %	Est. ponc.	Var.	Var. est.	Var. (var. est.)	EQM (var. est.)
Moyenne de Y	Données complètes	93,6	468,97	461,88	29 301,77	29 352,04
	EAS	31,1	259,46	303,46	10 228,40	12 164,74
	FX	93,7	473,86	474,21	30 408,95	30 409,07
	HM	93,4	476,03	465,53	29 406,61	29 516,85
Ordonnée à l'origine	Données complètes	93,0	451,46	432,74	14 955,20	15 305,73
	EAS	31,5	275,22	311,36	8 134,04	9 440,57
	FX	93,2	456,08	444,88	15 539,21	15 664,64
	HM	92,3	457,48	436,25	14 941,00	15 391,75
Pente	Données complètes	93,1	10,09	0,99	0,09	0,10
	EAS	59,0	7,72	1,67	0,35	0,36
	FX	93,4	10,10	1,03	0,98	0,10
	HM	93,3	10,10	1,03	0,96	0,10

Méthode	Conv. IC à 95 %	Est. ponc.	Var.	Var. est.	Var. (var. est.)	EQM (var. est.)	
Moyenne de Y	Données complètes	94,2	2,0	544,91	527,31	31 626,19	31 936,07
	EAS	38,0	45,8	327,79	360,74	11 927,97	13 013,35
	FX	94,8	2,4	554,09	579,92	37 474,82	38 141,70
Ordonnée à l'origine	HM	94,5	2,3	551,02	553,16	34 056,39	34 060,99
	Données complètes	93,0	2,4	529,51	499,73	18 543,13	19 430,21
	EAS	39,5	46,8	340,09	365,50	9 351,15	9 996,99
Pente	Données complètes	93,3	10,1	1,24	1,15	0,14	0,15
	FX	94,5	10,1	1,45	1,44	0,18	0,18
	HM	95,7	10,1	1,53	1,65	0,29	0,30

Propriétés des procédures d'imputation lorsque les coefficients de régression de population sont reliées à la variable d'enquête d'intérêt

Tableau 2

Le tableau 2 montre les résultats de 1 000 répétitions des trois stratégies d'imputation décrites au tableau 1. La ligne supplémentaire annotée « Données complètes » donne les résultats en utilisant les données pour toutes les unités échantillonnées, c'est-à-dire en supposant qu'aucune unité pour laquelle $I_{hij} = 1$ n'a $R_{hij} = 0$. La colonne étiquetée « Conv. IC à 95 % » contient le pourcentage des 1 000 intervalles de confiance simulés qui contiennent le paramètre de population. La colonne étiquetée « Est. ponc. » contient les moyennes des 1 000 estimations ponctuelles de \bar{Q} . La colonne étiquetée « Var. » contient les variances des 1 000 estimations ponctuelles de \bar{Q} . La colonne étiquetée « Var. est. » contient les moyennes sur les 1 000 répétitions des variances estimées des estimations ponctuelles. Les colonnes étiquetées « Var(var. est.) » et « EQM(var. est.) » donnent la variance et l'erreur quadratique moyenne des 1 000 variances estimées.

Toutes les imputations sont tirées d'après les lois bayésiennes prédictives à posteriori appropriées. Premièrement, nous sélectionnons les paramètres des modèles d'imputation à partir des lois a posteriori sachant les composantes des données observées, (Z, X, X^{obs}, I, R) , qui sont incluses dans les modèles. Deuxièmement, nous sélectionnons les valeurs des données manquantes à partir des lois données au tableau 1. Nous utilisons des lois a priori diffusées pour tous les paramètres. Pour la stratégie HM, nous tirons les valeurs des paramètres en utilisant un échantillonneur de Gibbs (Gelfand et Smith 1990). Nous exécutons l'échantillonneur pendant une période de rodage pour obtenir la convergence approximative, puis nous utilisons chaque dixième tirage pour les imputations. Enfin, nous utilisons $M = 5$ imputations tirées indépendamment dans chaque ensemble de données pour chaque stratégie.

3.1 Simulation A : Illustration de la non-prise en compte des caractéristiques pertinentes du plan

Dans cette simulation, nous générons une population dans laquelle la distribution de Y diffère selon la strate et la grappe. Nous l'appelons « Population 1 ». Plus précisément, pour l'unité j dans la strate h et la grappe c, nous construisons la valeur de population de Y_{hij} d'après

Étiquette	Modèle d'imputation pour X_{hij} manquante
EAS	$N(\beta_0 + \beta_1 X_{hij}, \sigma^2)$
FX	$N(\beta_0h + \beta_1h X_{hij} + \omega_{hc}, \sigma_h^2)$
HM	$N(\beta_0h + \beta_1h X_{hij} + \omega_{hc}, \sigma_h^2), \omega_{hc} \sim N(0, \tau^2)$

Stratégies d'imputation

Tableau 1

$$Y_{hij} = 10 X_{hij} + \beta_0h + \omega_{hc} + \epsilon_{hij} \quad (5)$$

Dans la troisième population, Y n'est relié ni aux indicateurs de strate ni aux indicateurs de grappe. Nous utilisons la première population pour démontrer qu'il importe d'inclure toutes les variables du plan pertinentes, et les deuxième et troisième populations, pour examiner l'effet de l'inclusion de variables du plan non pertinentes. Les populations simulées sont stylisées afin d'illustrer l'importance de la modélisation du plan de sondage; par conséquent, la grandeur des biais/inéficacités n'est pas nécessairement généralisable à d'autres conditions.

Chaque population est divisée en cinq strates de taille égale comprenant chacune $N_h = 200$ grappes, pour $h = 1, \dots, 5$. Chaque grappe c dans la strate h comprend N_{hc}^{hc} unités. Dans chaque strate, 10 grappes ont $N_{hc} = 300$, 20 grappes ont $N_{hc} = 200$, 60 grappes ont $N_{hc} = 100$, 60 grappes ont $N_{hc} = 75$, et cinquante grappes ont $N_{hc} = 50$. Nous faisons varier les tailles de grappe afin de grossir les effets du plan lors du tirage d'échantillons en grappes à plusieurs degrés. Pour chaque population cible, il existe deux variables d'enquête, X et Y . Dans les trois populations, par souci de simplicité, nous générons chaque X_{hcf} , où l'indice f indique une unité dans la strate c et la grappe h , à partir de $X_{hcf} \sim N(0, 10^2)$. Pour générer Y , nous utilisons différentes méthodes pour chaque population, comme nous le décrirons aux sections qui suivent.

Nous échantillonnons aléatoirement les unités à partir de chaque population en utilisant un plan d'échantillonnage en grappes à plusieurs degrés. Pour commencer, nous tirons un échantillon aléatoire simple de $n_1 = 40$ grappes à partir de la strate 1, $n_2 = 20$ grappes à partir de la strate 2, $n_3 = 30$ grappes à partir de la strate 3, $n_4 = 10$ à partir de la strate 4 et $n_5 = 15$ grappes à partir de la strate 5. Les tailles des échantillons en grappes varient selon la strate afin de grossir les effets de plan comparativement à l'échantillonnage uniforme. Puis, nous tirons un échantillon aléatoire simple de 20 unités dans chaque grappe échantillonnée. Donc, nous obtenons 2 300 unités pour lesquelles $I_{hcf} = 1$.

Dans chaque population, les paramètres estimés sont $\bar{Q} = \bar{Y}$, la moyenne de population de Y , et les coefficients de population de Y sur X . L'estimateur en données complètes de \bar{Y} est l'estimateur sans biais fondé sur le plan de sondage habituel,

$$\bar{y}_{hc} = N_{hc} \bar{y}_{hc} \quad \text{ou } \bar{y}_{hc} \text{ est le total estimé dans la grappe } hc. \\ \text{L'estimateur en données complètes de la variance de } \bar{y}_{hc} \text{ est}$$

$$q = \frac{1}{100\,000} \left(\sum_{h=1}^5 \sum_{c=1}^{200} \bar{y}_{hc}^2 \right) - \left(\sum_{h=1}^5 \frac{1}{200} \sum_{c=1}^{200} \bar{y}_{hc} \right)^2$$

$$n = \frac{1}{100\,000^2} \left(\sum_{h=1}^5 \sum_{c=1}^{200} \bar{y}_{hc}^2 \right) - \left(\sum_{h=1}^5 \frac{1}{200} \sum_{c=1}^{200} \bar{y}_{hc} \right)^2$$

de données multi-imputés dans toutes les simulations. Pour chaque unité, la variable de réponse binaire, R_{hcf} , est tirée à partir d'une loi de Bernoulli :

$$\Pr(R_{hcf} = 1) = \frac{\exp(-0,847 - 0,1X_{hcf})}{1 + \exp(-0,847 - 0,1X_{hcf})} \quad (4)$$

Ici, $R_{hcf} = 1$ signifie que la valeur de Y manque pour l'unité hcf . L'équation 4 implique que X_{hcf} manque au hasard (Rubin 1976). Nous pouvons ignorer le mécanisme de création des données manquantes à condition que les imputations pour ces données soient conditionnelles à X . Délibérément, nous ne permettons pas que l'absence de données dépende de l'appartenance à la strate ou à la grappe, afin d'illustrer que le biais peut être dû au fait de ne pas tenir compte du plan de sondage, même si le mécanisme de création de données manquantes ignorable ne dépend pas du plan d'échantillonnage. Naturellement, si le plan d'échantillonnage est relié au fait que des données manquent, comme cela est le cas dans de nombreux ensembles de données réels, il faut introduire les contraintes du plan d'échantillonnage afin que le mécanisme de création des données manquantes soit ignorable.

Nous examinons trois stratégies d'imputation de X_{hcf} s'appuyant sur différentes utilisations de l'information sur le plan de sondage. Ces stratégies sont résumées au tableau 1. La première, dénotée EAS, omet entièrement de tenir compte du plan d'échantillonnage. La deuxième, dénotée FX, intègre la stratification et la mise en grappes grâce à l'utilisation d'effets fixes pour chaque grappe dans la strate. La troisième stratégie, dénotée HM, consiste à utiliser des modèles normaux à effets aléatoires dans lesquels sont intégrées la stratification et la mise en grappes. Pour EAS, un modèle est ajusté à l'ensemble de données complet. Pour FX et HM, les modèles sont ajustés séparément à chaque strate. Les trois stratégies comportent la régression sur X , parce que cette variable fait partie du mécanisme de création des données manquantes; ne pas conditionner à X violerait l'ignorabilité et causerait un biais.

s'appuie sur des modèles hiérarchiques où i) les effets de la mise en grappes sont intégrés en utilisant des effets aléatoires et ii) les effets de la stratification sont intégrés en utilisant des effets fixes. Les simulations montrent que tenir compte du plan de sondage de cette façon peut réduire le biais. Elles illustrent aussi le fait qu'introduire des caractéristiques du plan qui ne sont pas reliées aux variables de l'enquête peut donner lieu à des inférences inefficaces, mais prudentes, comparativement à celles faites d'après des modèles dans lesquels ce genre de caractéristiques ne sont pas intégrées comme contraintes, à condition que les modèles incluent les variables explicatives requises pour que l'hypothèse selon laquelle les données manquent au hasard (Rubin 1976) soit plausible. Nous démontrons la première approche d'intégration des caractéristiques du plan en imputant des données manquantes dans le cas de la National Health and Nutrition Examination Survey selon une méthode de régression séquentielle.

2. Inférences d'après des ensembles de données multi-imputés

Afin de décrire la construction d'ensembles de données multi-imputés et les inférences d'après ces derniers, nous utilisons la notation de Rubin (1987). Pour une population finie de taille N , soit $I_j = 1$ si l'unité j est sélectionnée dans l'enquête originale, et $I_j = 0$ autrement, où $j = 1, 2, \dots, N$. Soit $I = (I_1, \dots, I_N)$. Soit n la taille d'échantillon obtenue au moyen d'un plan d'échantillonnage complexe. Pour simplifier la notation, supposons qu'une seule variable de l'enquête est sujette à la non-réponse. Soit $R_j = 1$, si l'unité j répond à l'enquête originale, et $R_j = 0$, autrement. La notation peut être étendue afin de traiter la non-réponse partielle multivariée, mais ce genre de complication n'est pas nécessaire aux fins de notre exposé. Soit X la matrice de données d'enquête de dimensions $N \times p$ pour toutes les unités de la population. Soit $X_{inc} = (X_{inc}^{obs}, X_{inc}^{mis})$ la matrice de données d'enquête de dimensions $n \times p$ pour les unités pour lesquelles $I_j = 1$; X_{inc}^{obs} est la portion de X_{inc} qui est observée, et X_{inc}^{mis} est la portion de X_{inc} qui manque à cause de la non-réponse. Soit Z la matrice de variables du plan de dimensions $N \times d$ pour les N unités de la population, par exemple, des indicateurs de strates ou de grappes ou des mesures de taille. Nous supposons que ce genre d'information sur le plan est connue au moins approximativement, par exemple d'après les dossiers du recensement ou les bases de sondage.

$$D^{(i)} = (Z, Y, X_{inc}^{obs}, Y^{(i)}, X_{inc}^{mis}, I, R).$$

Pour obtenir M ensembles de données complets, Ces tirages sont répétées indépendamment $I = 1, \dots, M$ fois bayésienne prédictive à posteriori de $(X_{inc}^{mis}, Y^{(i)}, I, R)$.

L'analyste peut alors utiliser \bar{q}_M pour estimer \bar{Q} et $T_M = (1 + \frac{1}{M})b_M + \bar{u}_M$ pour estimer la variance de \bar{q}_M . Quand n et M sont grands, les inférences pour le scalaire \bar{Q} peuvent être fondées sur des lois normales, de sorte qu'un intervalle de confiance à $(1 - \alpha)\%$ pour \bar{Q} est $\bar{q}_M \pm z(\alpha/2)\sqrt{T_M}$. Pour une valeur modérée de M , les inférences peuvent être fondées sur des lois t avec $\nu_M = (M - 1)(1 + r_M^{-1})^\gamma$ degrés de liberté, où $r_M = (1 + M^{-1})b_M/\bar{u}_M$, et γ est sorte qu'un intervalle de confiance à $(1 - \alpha)\%$ pour \bar{Q} est $\bar{q}_M \pm t_{\nu_M}(\alpha/2)\sqrt{T_M}$. Des perfectionnements de ces règles plusieurs auteurs, y compris Li, Meng et Rubin (1991a), Raghunathan et Rubin (1991b), Meng et Rubin (1991b), Raghunathan et Siscovick (1996), ainsi que Barnard et Rubin (1999).

3. Simulations illustratives

À la présente section, nous utilisons des simulations pour illustrer les biais/inefficacités associées à l'intégration des caractéristiques du plan dans les modèles d'imputation. Nous simulons trois populations cibles de $N = 100\,000$ unités, qui sont stratifiées et mises en grappes dans les strates. Dans la première population, Y dépend à la fois des effets de strate et de grappe. Dans la deuxième population, Y dépend des effets de strate, mais non des effets de grappe.

L'importance de la modélisation du plan d'échantillonnage dans l'imputation pour les données manquantes

Jerome P. Reiter, Trivellore E. Raghunathan et Satkartar K. Kinney

Résumé

La théorie de l'imputation multiple pour traiter les données manquantes exige que l'imputation soit faite conditionnellement du plan d'échantillonnage. Cependant, comme la plupart des logiciels standard utilisés pour l'imputation multiple fondée sur un modèle reposent sur l'hypothèse d'un échantillonnage aléatoire simple, de nombreux praticiens sont portés à ne pas tenir compte des caractéristiques des plans d'échantillonnage complexes, comme la stratification et la mise en grappes, dans leurs imputations. Or, la théorie prédit que l'analyse d'ensembles de données soumis de telle façon à une imputation multiple peut produire des estimations biaisées du point de vue du plan de sondage. Dans le présent article, nous montrons au moyen de simulations que i) le biais peut être important si les caractéristiques du plan sont reliées aux variables d'intérêt et que ii) le biais peu être réduit en tenant compte de l'effet des caractéristiques du plan dans les modèles d'imputation. Les simulations montrent aussi que l'introduction de caractéristiques non pertinentes du plan comme contraintes dans les modèles d'imputation peut donner lieu à des inférences conservatrices, à condition que les modèles contiennent aussi des variables explicatives pertinentes. Ces résultats portent à formuler la prescription qui suit à l'intention des imputeurs : le moyen le plus sûr de procéder consiste à inclure les variables du plan de sondage dans une démonstration d'une approche simple d'intégration des caractéristiques d'un plan de sondage complexe qui peut être suivie en utilisant certains logiciels standard pour créer des imputations multiples.

Mots clés : Plan de sondage complexe; imputation multiple; non-réponse; enquêtes.

1. Introduction

En général, dans les grandes enquêtes, les unités échantillonées ne répondent pas toutes complètement au questionnaire. Certaines n'y répondent pas du tout et d'autres ne répondent qu'à certaines questions. Une approche pour traiter ce genre de non-réponse est l'imputation multiple des données manquantes (Rubin 1987). Elle a été utilisée, par exemple, dans le *Fatality Analysis Reporting System* (Heitjan et Little 1991), la *Consumer Expenditures Survey* (Raghunathan et Paulin 1998), la *National Health and Nutrition Examination Survey* (Schafer, Ezzati-Rice, Johnson, Khare, Little et Rubin 1998), la *Survey of Consumer Finances* (Kienickell 1998) et la *National Health Interview Survey* (Schenker, Raghunathan, Chiu, Makuc, Zhang et Cohen 2005). L'imputation multiple a également été proposée pour assurer la protection des renseignements personnels dans les fichiers de données à grande diffusion (Rubin 1993; Little 1993; Raghunathan, Reiter et Rubin 2003; Reiter 2004, 2005). Pour une revue d'autres applications, voir Rubin (1996), ainsi que Barnard et Meng (1999).

En théorie, lors de l'établissement de méthodes d'imputation d'après des ensembles de données ayant subi une imputation multiple, cette dernière est rendue conditionnelle au plan d'échantillonnage (Rubin 1987). Toutefois, les imputeurs tiennent rarement compte des caractéristiques des

plans d'échantillonnage complexes, comme la stratification et la mise en grappes, lorsqu'ils utilisent les logiciels disponibles pour construire des modèles d'imputation. Ils se servent plutôt de modèles normaux ou de modèles de localisation multivariés (par exemple, le logiciel NORM rédigé par Joe Schafer), ou de modèles de régression séquentielle (Raghunathan, Lepkowski, van Hoewyk et Solenberger 2001). Bien que ces méthodes puissent être modifiées afin d'intégrer les caractéristiques du plan, cela se fait rarement.

L'objectif du présent article est double. En premier lieu, nous illustrons le biais qui peut se produire lorsque les imputeurs omettent de tenir compte des caractéristiques du plan de sondage complexe dans les modèles d'imputation. Pour cela, nous simulons une imputation multiple dans des échantillons à deux degrés, stratifiés et mis en grappes. Les simulations indiquent que le biais peut être important, même si l'on applique des estimateurs fondés sur le plan de sondage à des ensembles de données soumis à l'imputation multiple ne présentant qu'une quantité modérée de données manquantes. En deuxième lieu, nous proposons deux approches simples en vue de tenir compte des caractéristiques du plan dans les modèles d'imputation. La première, qui est relativement facile à mettre en œuvre, comprend des variables nominales pour les effets de strate, ou de grappes dans les modèles d'imputation. La deuxième, qui requiert des calculs plus compliqués que la première,

Kott, P.S. (1990). The design consistent regression estimator and its conditional variance. *Journal of Statistical Planning and Inference*, 24, 287-296.

Kott, P.S. (1994). A note on handling nonresponse in surveys. *Journal of the American Statistical Association*, 89, 693-696.

Kott, P.S. (2004a). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 48, 263-277.

Kott, P.S. (2004b). Commentaire. *Techniques d'enquête*, 30, 28-29.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 139-157.

Lundström, S., et Sämäl, C.-E. (1999). Calibration as a standard method for the treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

Sämäl, C.-E., Swensson, B. et Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of a finite population total. *Biometrika*, 76, 527-537.

Sämäl, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

un ensemble de poids. Notons que, puisque les poids de calage peuvent être négatifs dans le cas de la méthode linéaire, celle-ci pourrait produire un ensemble que la méthode par ajustement proportionnel itératif généralisé ne pourrait pas produire. La méthode linéaire rééchelonne effectivement les a_i , c'est-à-dire la valeur de chaque élément dans le même groupe, d'une quantité fixe. Donc, elle pourrait ne pas produire des poids étonnamment petits ou étonnamment grands lorsque la dimension de \mathbf{x}_k est faible comparativement à la taille d'échantillon.

9.3 Calage de l'échantillon et calage pour la non-réponse

À la section précédente, nous avons indiqué qu'il est possible que les composantes de \mathbf{h}_k dans l'équation (13), c'est-à-dire le modèle de réponse quasi aléatoire, soient inconnues avant le recensement. Lorsqu'on utilise un tel \mathbf{h}_k

dans le calage, il n'est peut-être plus raisonnable d'affirmer que l'estimateur $t_{y, CAL}$ résultant est sans biais par rapport au modèle prédicatif. Cela est particulièrement ennuyeux lorsque la non-réponse est modérée, comparativement à la taille de l'échantillon. Une idée intéressante consiste à faire le calage en deux phases. La première phase, le calage de l'échantillon, consiste à corriger pour la différence entre T_k et $\sum_F a_k \mathbf{x}_k$, et ne comprendrait aucune composante de \mathbf{h}_k inconnue au moment de l'échantillonnage. La deuxième phase, le calage pour la non-réponse, corrige pour la différence entre $\sum_F a_k \mathbf{x}_k$ et $\sum_S a_k \mathbf{x}_k$ et comprendrait uniquement les variables composantes disponibles après que le sous-échantillon de répondants soit recensé.

Une analyse plus approfondie de cette idée doit être reportée à une autre occasion.

9.4 Travaux avec le NASS

Le National Agricultural Statistics Service (NASS) a utilisé des variantes de l'approche de Fuller et coll. (1994) pour traiter le sous-dénombrement au Recensement de l'agriculture de 2002 (voir Fetter et Kott 2003) et pour la correction d'une enquête sur l'économie agricole avec non-réponse importante, de façon à faire concorder les totaux à ceux d'enquêtes plus fiables (voir Crouse et Kott 2004). Dans cette approche, $f(\cdot)$ est de la forme :

$$f(\mathbf{x}_k \phi) = \begin{cases} L & \text{si } \mathbf{x}_k \phi > L \\ L \leq \mathbf{x}_k \phi \leq U & \text{si } \mathbf{x}_k \phi \leq U \\ U & \text{si } \mathbf{x}_k \phi > U \end{cases} \quad (20)$$

qui tronque le calage linéaire à des valeurs préétablies, L et U , pour contrôler l'importance de l'ajustement des poids. Notons que, quand $f(\cdot) = U$ ou L , $f'(\cdot) = 0$. Contrairement à l'ajustement par calage de l'équation (19), $f(\cdot)$ de l'équation (20) n'est pas dérivable deux fois aux valeurs L ou U . Cela ne cause pas de problème en pratique.

Bibliographie

Berry, C.C., Flatt, S.W. et Pierce, J.P. (1996). Correcting unit nonresponse via nonresponse modeling and raking in the California Tobacco Survey. *Journal of Official Statistics*, 12, 349-363.

Crouse, C., et Kott, P.S. (2004). Evaluation alternative calibration schemes for an economic survey with large nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Deming, W.E., et Stephan, F.F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal total are known. *Annals of Mathematical Statistics*, 11, 427-444.

Dennett, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 17-27.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Estévez, V.M., et Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16, 379-399.

Fetter, M.J., et Kott, P.S. (2003). Developing a coverage adjustment strategy for the 2002 Census of Agriculture. Présenté à 2003 Federal Committee on Statistical Methodology Research Conference, http://www.fcsm.gov/03papers/fetter_kott.pdf.

Folsom, R.E., et Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 598-603.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.

Fuller, W.A., Louhghagh, M.M., et Baker, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey 1987-88. *Techniques d'enquête*, 20, 79-89.

Kim, J.K. (2004). Efficient nonresponse weighting adjustment using estimated response probability. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Statistique Canada, N° 12-001-XPB au catalogue

le traitement des erreurs de couverture.

En présence de non-réponse totale, nombreux sont ceux qui ont essayé d'estimer les probabilités de réponse individuelles, $p_i = 1/f(\mathbf{h}_i, \phi)$, directement. Cette méthode requiert de l'information sur \mathbf{h}_i pour chaque unité échantillonnée, qu'elle réponde à l'enquête ou non, mais \mathbf{h}_i ne doit pas avoir la même dimension que \mathbf{x}_i . La méthode d'ajustement directe n'est généralement pas disponible pour

9.1 Estimation explicite d'un modèle de réponse

9. Discussion

Un deuxième ensemble de 1 600 simulations (non utilisables, présentées) ont été exécutées en utilisant la même population et un plan d'échantillonnage stratifié, mais en dominant à chaque élément échantillonné 70 % de chances de faire partie de l'échantillon de répondants (la taille moyenne de l'échantillon de répondants était d'environ 89,8). Dans cet ensemble de simulations, les deux estimateurs de T_y sont convergents par rapport au plan d'échantillonnage (randomisation) sous le modèle de réponse. Par conséquent, il n'est pas étonnant que les moyennes empiriques de t_{y_NIT} et de t_{y_EXP} soient presque identiques (écart d'au plus 0,01 %), comme le sont leurs erreurs quadratiques moyennes empiriques (écart d'au plus 1 %). Les moyennes empiriques de chaque paire d'estimateurs de variance (par exemple var_{ST2} pour t_{y_LIN} et t_{y_EXP}) étaient aussi très proches (écart d'au plus 1 %). Le biais relatif de l'estimateur var_{ST2} ajusté (comparativement à $var_{ST2(e)}$) était de -1,3 % lors de l'estimation de la variance de t_{y_LIN} et de -2,2 % lors de l'estimation de la variance de t_{y_EXP} . Le biais relatif des variances par linéarisation non corrigées était de -9,0 % et de -10,3 %, respectivement. Les biais relatifs des deux estimateurs jacksonnité était de 3,6 %.

Qu'il soit préférable ou non d'intégrer le modèle de réponse correcte dans l'estimateur de calage, si on le fait, alors les estimateurs de variance discutés à la section précédente, peut-être avec l'estimateur par linéarisation corrigé comme il est suggéré à la présente section, semblent

Lors de l'utilisation de $V_{\Sigma(2)}^{\text{L2}}$ comme approximation efficace de l'EOMF, l'erreur quadratique moyenne empirique de t_{EXP} , qui intégrait le modèle de réponse correcte, était plus de 13 % plus importante que celle de t_{LIN} , qui n'intégrait pas ce modèle. Toutefois, il ne convient pas de faire de grandes généralisations à partir d'un ensemble de données comportant deux variables de calage seulement. Voir Crouse et Koit (2004) pour un ensemble différent de

réponse ».

Un autre moyen d'intégrer les valeurs ajustées de $f(\mathbf{u})$ dans l'estimation fondée sur la méthodologie développée dans le texte est décrit ci-après. Répartir les valeurs ajustées en P groupes d'après leur taille, où P est de nouveau la dimension de \mathbf{x}_i , et soit \mathbf{d}^k , un vecteur ligne de variables indicatrices pour les P cellules. En fixant chaque $w_i = a^k [1 + (T_i - \sum a^j x_i^j) / (\sum a^j \mathbf{d}^j \mathbf{d}^k)]$, on calcule un ensemble de poids pour le sous-échantillon de répondants qui, contrairement à $\{w_i^{ADV}\}$ ci-dessus, satisfait l'équation de calage pour l'échantillon de répondants. Étant donné la structure de \mathbf{d}^k , cette méthode linéaire produit le même ensemble de poids de calage que celui que donnerait l'ajustement de $w_i = a^k \exp(\mathbf{d}^k \mathbf{p})$ — si les deux produisaient

$$w_{k-\text{adj}} = \sum_{c \in \mathcal{C}} \sum_{m \in \mathcal{M}} w_m^{s(c)} w_k^{s(c)} = \sum_{c \in \mathcal{C}} \sum_{m \in \mathcal{M}} w_m^{s(c)} w_k^{s(c)}$$

Afin de contrôler l'importance de la repondération due à la non-réponse, Little (1986) a recommandé que l'on estime \mathbf{p} explicitement, puis qu'on divise l'échantillon en C groupes mutuellement exclusifs en se fondant sur les tailles des valeurs ajustées de $f(\mathbf{h}^*)$. On calcule ensuite le poids corrigé pour chaque élément k dans le groupe c comme dans le cas de la poststratification :

9.2 Groupes à homogénéité de réponse

Un compromis raisonnable consiste à choisir la forme de $f(\cdot)$ et de \mathbf{h}_k en modélisant le comportement de réponse de l'échantillon complet, puis à estimer le paramètre $f(\cdot)$ implicitement par calage. Ce compromis permet aussi de contourner une faiblesse frappante de l'utilisation de la pondération par calage pour corriger de la non-réponse (ainsi que pour les erreurs de couverture). Les choix de $f(\cdot)$ et \mathbf{h}_k sont motivés principalement par la vraisemblance et la commodité et non par une analyse statistique des données.

de répondre, ainsi, $y^a = a^a + f + f(b^a b^a)$, et \mathbf{b} est un estimateur direct convergent pour le paramètre du modèle de quasi-randornisation, ϕ . Cela ne sous-entend pas que l'estimation directe du modèle de réponse fondée sur une $f(\cdot)$ et un \mathbf{h}^a donnés est moins efficace que le calage analogue lorsque \mathbf{h}^a a la même dimension que \mathbf{x}^a . Voir Kim (2004) pour une suggestion du contraire. Néanmoins, la commodité qu'offre l'intégration de la correction pour la non-réponse dans le calage est séduisante, lorsque les estimations de

Pour $f'(\mathbf{x}_j^{\text{LIN}})$, $f'(\mathbf{x}_j^{\text{EXP}}) = \exp(\mathbf{x}_j^{\text{EXP}}) = \exp(\mathbf{x}_j^{\text{LIN}} \mathbf{p}) = 1$; pour $f'(\mathbf{x}_j^{\text{EXP}})$ et $f'(\mathbf{x}_j^{\text{LIN}}) = 1/p_j$.

Le tableau 1 donne les moyennes empiriques (la moyenne sur les 1 600 simulations) des deux estimateurs pour T_y normalisés de sorte que $T_y = 100$. Bien que tous deux soient pratiquement sans biais, $f'_{y, \text{LIN}}$ diffère significativement de 0,05; ce n'est pas le cas pour $f'_{y, \text{EXP}}$. Cela n'est pas étonnant, parce que seul le dernier est fondé sur le modèle de réponse correcte. Pour chaque estimateur, les estimateurs de variance et les erreurs quadratiques moyennes empiriques ont été normalisés de sorte que les moyennes empiriques des $V_{ST2(e)}$ respectifs soient égales à 100. Aucun $V_{ST2(e)}$ n'avait une moyenne empirique significativement différente de l'erreur quadratique moyenne empirique (EQME) de l'estimateur associé. Ce résultat était un peu décevant. Il semble que, bien que $f'_{y, \text{LIN}}$ ait un biais empirique significatif, celui-ci était une composante tellement faible de l'erreur quadratique moyenne de l'estimateur que la différence entre l'EQME de cet estimateur et la moyenne empirique de $V_{ST2(e)}$ n'était pas significative.

Les $V_{ST2(e)}$ ont été choisis comme valeurs de référence pour le tableau plutôt que les erreurs quadratiques moyennes empiriques, parce que chaque $V_{ST2(e)}$ avait environ la moitié de l'erreur-type empirique de l'EQME correspondant (qui, elle-même, correspondait à la moyenne des 1 600 carrés des écarts) et était corrélée plus fortement avec les estimateurs $V_{ST2(e)}$ n'était pas significative.

Les $V_{ST2(e)}$ ajusté pour $f'_{y, \text{LIN}}$ ainsi que $f'_{y, \text{EXP}}$ demeuraient enache d'un biais par défaut, tandis que le V_j présentait un biais par excès d'une valeur légèrement plus faible. Bien que ces biais soient significatifs, ils étaient raisonnablement petits (de 4,5 à 11,2 %) et donnent à penser que les estimateurs de variance étaient peut-être effectivement asymptotiquement sans biais, comme nous l'avons démontré théoriquement aux sections précédentes.

Tableau 1

Moyennes empiriques des estimateurs basées sur 1 600 simulations*

Estimateurs pour T_y ($T_y = 100$)		Moyenne empirique (erreur-type)		Valeur t (test de signification bilatéral)	
$f'_{y, \text{LIN}}$	99,84 (0,06)	100,04 (0,06)	$f'_{y, \text{EXP}}$	0,58 (0,56)	différent de T_y
Estimateurs de variance pour $f'_{y, \text{LIN}}$ ($E_{\text{EMP}}(V_{ST2(e)}) = 100$)		83,39 (1,53)	95,33 (1,80)	104,69 (2,28)	V_j
V_{ST2}	83,39 (1,53)	95,33 (1,80)	$V_{ST2(\text{ajusté})}$	99,35	EQME
V_j	104,69 (2,28)	99,35	Estimateurs de variance pour $f'_{y, \text{EXP}}$ ($E_{\text{EMP}}(V_{ST2(e)}) = 100$)	73,12 (1,54)	V_{ST2}
V_{ST2}	73,12 (1,54)	88,79 (1,98)	$V_{ST2(\text{ajusté})}$	107,00 (2,73)	V_j
V_j	107,00 (2,73)	101,21	EQME	0,33 (0,74)	Autres statistiques
relvar ($V_{ST2(e)}^{\text{LIN}}$)		0,051	relvar ($V_{ST2(e)}^{\text{EXP}}$)	—	
$(V_{ST2(e)}^{\text{LIN}} - V_{ST2(e)}^{\text{EXP}})$		—0,1340 (0,010)	$(E_{\text{EMP}}(V_{ST2(e)}^{\text{EXP}}))$	—13,87 (< 0,0001)	

* Dans quatre simulations supplémentaires, la convergence n'a pas été atteinte en dix itérations pour $f'_{y, \text{EXP}}$. Ces simulations ont été exclues de l'analyse.

8. Un petit exemple empirique

Puisque les poids de rééchantillonnage jackknife exprimés par l'équation (18) sont nouveaux, il est prudent de chercher à savoir s'ils fonctionnent effectivement avec des données réelles. Pour ce faire, nous avons pris les données MJ281 de Samdal, Swensson et Wetman (1992) et les avons répétées 20 fois (de sorte que $N = 5\,620$). Par échantillonnage aléatoire simple stratifié, nous avons sélectionné 16 unités dans chacune des huit strates de taille inégale. La variable RMT85 a servi de y_k et la variable $P75$, de x_k dans $\mathbf{x}_k = (1, x_k)$. À chacune des 128 unités échantillonnées, nous avons attribué une probabilité d'être présente dans le sous-échantillon de répondants, S , qui diminuait avec la taille de x_k ; en particulier, $P_k = \exp(-0,35\,x_k/M^x)$, où M^x était la moyenne de population de x_k . Dans les 1 600 simulations, la taille de S variait de 78 à 110, avec une moyenne d'environ 93,8.

Le total T_y a été estimé de deux façons, avec $t_{y,NT} = \sum_{k \in (ex)} \mathbf{x}_k \mathbf{b}^T y_k$ et avec $t_{y,EXP} = \sum_{k \in (ex)} a_k \exp(\mathbf{x}_k \mathbf{b}^T y_k)$, où \mathbf{b} et $\mathbf{b}^{(ex)}$ étaient, respectivement, sélectionnés de sorte que l'équation de calage soit vérifiée. Le premier était un estimateur GREG, tandis que le second était un estimateur par ajustement proportionnel itératif généralisé. Les deux estimateurs étaient sans biais sous le modèle prédictif sous-entendu ($y_k = \mathbf{x}_k \beta + \varepsilon_k$), mais seul $t_{y,EXP}$ était convergent dans des conditions de randomisation sous le modèle de réponses correcte. L'estimateur GREG supposait implicitement que $P_k = 1/(\phi_{(NT)}^0 + \phi_{(NT)}^1 \mathbf{x}_k)$ pour $\phi_{(NT)}^0$ et $\phi_{(NT)}^1$ inconnus.

La petite taille de l'échantillon comparativement à la population de chaque strate a permis d'ignorer la correction pour population finie dans l'estimation de la variance/erreur quadratique moyenne (appelée dans la suite « estimation de la variance »). Nous avons estimé les variances en utilisant l'estimateur par linéarisation, $_{VST2}$, dans l'équation (10) t_y défini par l'équation (16) et t_y le jackknife proposé, avec n_k dans l'équation (11) avec les poids de rééchantillonnage définis par l'équation (18). Pour rendre le calcul du jackknife plus simple, les 16 sous-échantillons dans chaque strate ont été attribués aléatoirement à l'une de quatre grappes, de sorte que 32 répliques jackknife seulement ont dû être calculées.

Aux fins de comparaison, une meilleure version de l'estimateur de variance par linéarisation, dénotée $_{VST2(e)}$, a également été calculée avec n_k remplacé par $e_k = y_k - \phi$, où $\phi = f(\mathbf{x}^j(\phi)/d^j \mathbf{x}^j(\phi))^{-1} \sum_{j=1}^J f(\mathbf{x}^j(\phi)/d^j \mathbf{x}^j(\phi))$, e_k est rarement connu, mais le calcul de $_{VST2(e)}$ est utile ici pour les comparaisons. Il convient de souligner que les calculs de n_k et e_k diffèrent légèrement selon que l'on voulait calculer l'estimateur de la variance pour $t_{y,LIN}$ ou pour $t_{y,EXP}$.

aléatoire, de l'échantillonnage à lieu conceptuellement avant population associée à la base de sondage est un échantillon de Poisson provenant d'une population complète hypothétique pour laquelle le vecteur T_y doit être connu. La population de la base de sondage devient F_y tandis que la population complète hypothétique est U . Nous supposons que la probabilité que l'élément $k \in U$ soit dans F est modélisée correctement par l'équation (13). Si la première (de U à F) et la deuxième (de F à S) phases d'échantillonnage sont indépendantes, alors toute la théorie élaborée pour l'utilisation de la pondération par calage en vue de traiter la non-réponse peut être transposée au traitement des sous-décomposés.

Il convient de souligner que la correction de l'erreur de couverture par calage est une extension de la pratique bien connue de correction par poststratification souvent utilisée dans les enquêtes téléphoniques. Comme dans le cas partiel de la poststratification, il faut utiliser comme cible de calage pour U des quantités que l'on peut supposer dépourvues d'erreur ou ayant une erreur quadratique moyenne très faible comparativement aux estimateurs par calage. Folsom et Singh ont fait remarquer que le sur-décombrement (décombrement multiple) ou une combinaison de sous- et de sur-décombrement peuvent être traités en suivant leur méthode. La définition de P_k dans l'équation (13) devient le nombre prévu de fois que k est présent dans la base de sondage, nombre qui peut maintenant être supérieur à 1 à cause du décombrement multiple éventuel. Folsom et Singh proposent en outre de donner à $f(\cdot)$ la forme flexible :

$$f(\mathbf{x}_k \phi) = \frac{U(C - L) \exp(\mathbf{x}_k \phi) + L(U - C)}{U(C - L) \exp(\mathbf{x}_k \phi) + L(U - C)} \quad (19)$$

où $L > 0$, $1 < U < \infty$, et $L < C \leq U$ sont des constantes prédéterminées. Ils donnent à cette expression le nom de « modèle exponentiel général » ou « MEG ». Observons que, si $C = 1$, $U = \infty$, et $L = 0$, $P_k = 1/f(\mathbf{x}_k \phi) = \exp(-\mathbf{x}_k \phi)$. Similairement, si $C = 2$, $U = \infty$, et $L = 1$, $P_k = [1 + \exp(\mathbf{x}_k \phi)]^{-1}$; autrement dit, la probabilité de couverture (ou de réponse) est logistique. Les valeurs L et U servent de bornes sur l'ajustement par calage, $f(\cdot)$, tandis que $C = f(0)$ est effectivement son centre. Les auteurs ont rendu l'ajustement par calage dans le MEG encore plus souple en postulant trois classes d'unités d'échantillonnage, chacune ayant son propre ensemble de valeurs U , C et L . Ils ont proposé son utilisation pour la correction de l'erreur de couverture ainsi que la non-réponse totale.

$t_{y-CAL} - T_y = \sum_{k \in S} a_k^k f(\mathbf{h}_k \mathbf{q}) \gamma_k^k - \sum_{k \in U} \gamma_k^k$
 $= \sum_{k \in S} a_k^k f(\mathbf{h}_k \mathbf{q}) e_k^k - \sum_{k \in U} e_k^k$
 ou

$e_k^k = \gamma_k^k - \mathbf{x}_k^k \left(\sum_{j \in U} f(\mathbf{h}_j \phi) d_j \mathbf{h}_j^j \mathbf{x}_j^j \right)^{-1} \sum_{j \in U} f(\mathbf{h}_j \phi) d_j \mathbf{h}_j^j \mathbf{x}_j^j \gamma_j^j$
 et $d_j^j = 1/f(\mathbf{h}_j \phi)$. Les termes e_k^k sont de nouveau inconnus. Ils ont été conçus de sorte que $\sum_S a_k^k f(\mathbf{h}_k \phi) \mathbf{h}_k^k e_k^k = \mathbf{O}^p(N/\sqrt{n})$. En poursuivant :

$$\begin{aligned} t_{y-CAL} - T_y &= \sum_{k \in S} a_k^k f(\mathbf{h}_k \phi) e_k^k - \sum_{k \in U} e_k^k + \sum_{k \in S} a_k^k f(\mathbf{h}_k \phi) \mathbf{h}_k^k \mathbf{q} \{f(\mathbf{h}_k \mathbf{q})\} e_k^k \\ &= \sum_{k \in S} a_k^k f(\mathbf{h}_k \phi) e_k^k - \sum_{k \in U} e_k^k + \sum_{k \in S} a_k^k f(\mathbf{h}_k \phi) \mathbf{h}_k^k \mathbf{q} + \sum_{k \in S} a_k^k f(\mathbf{h}_k \phi) \mathbf{h}_k^k \mathbf{q} \{f(\mathbf{h}_k \mathbf{q})\} e_k^k \\ &= \sum_{k \in S} a_k^k f(\mathbf{h}_k \phi) \mathbf{h}_k^k \mathbf{q} \{f(\mathbf{h}_k \mathbf{q})\} e_k^k - \sum_{k \in U} e_k^k + \mathbf{O}^p(N/\sqrt{n}) \\ &= \sum_{k \in S} a_k^k f(\mathbf{h}_k \phi) \mathbf{h}_k^k \mathbf{q} \{f(\mathbf{h}_k \mathbf{q})\} e_k^k - \sum_{k \in U} e_k^k + \mathbf{O}^p(N/\sqrt{n}). \end{aligned} \quad (14)$$

Donc, t_{y-CAL} est convergent sous quasi-randomisation sous des contraintes faibles quand $t = \sum_S a_k^k f(\mathbf{h}_k \phi) \gamma_k^k$. l'est.

Pour estimer l'erreur quadratique moyenne sous quasi-randomisation de t_{y-CAL} (c'est-à-dire, l'erreur quadratique moyenne de l'estimateur dans les conditions de randomisation sous le modèle de réponses), nous commençons par noter que la probabilité que les éléments k et j , $k \neq j$, soient tous deux compris dans le sous-échantillon de répondants est $\pi_{kj}^* = \pi_k^* d_j^j$. Soit $\pi_k^* = \pi_k^* d_k^k$, et rappelons que $a_k^k = 1/\pi_k^k$ et $1/d_k^k = f(\mathbf{h}_k \phi)$. Partant de l'équation (14), nous voyons que l'erreur quadratique moyenne sous quasi-randomisation de t_{y-CAL} est approximativement

$$\begin{aligned} E_1[(t_{y-CAL} - T_y)^2] &\approx \sum_{k \in U} \sum_{j \in U} (\pi_{kj}^* - \pi_k^* \pi_j^*) (e_k^k / \pi_k^*) (e_j^j / \pi_j^*) \\ &= \sum_{k \in U} (1 - \pi_k^*) e_k^k / \pi_k^* + \sum_{k \in U} \sum_{j \in U} (\pi_{kj}^* - \pi_k^* \pi_j^*) (e_k^k / \pi_k^*) (e_j^j / \pi_j^*) \\ &= \sum_{k \in U} \sum_{j \in U} \left[\sum_{l \in S} a_l^l f(\mathbf{h}_l \phi) \mathbf{h}_l^l \mathbf{x}_l^l \right]^{-1} \sum_{l \in S} a_l^l f(\mathbf{h}_l \phi) \mathbf{h}_l^l \mathbf{x}_l^l \mathbf{h}_k^k \mathbf{h}_j^j \mathbf{x}_j^j \\ &= \sum_S (w_k^k w_j^j - w_k^k) t_{kj}^k \text{ avec} \end{aligned} \quad (16)$$

sert à la fois d'estimateur raisonnable de la variance fondée sur le modèle prédictif et de l'erreur quadratique moyenne sur le modèle de quasi-randomisation sous des contraintes faibles, puisque $w_k^k \approx 1/\pi_k^*$ et $t_{kj}^k \approx e_k^k$. Un proche parent du résidu d'échantillon non intuitif dans l'équation (16) est donné dans Folsom et Singh (2000). Voir

Kott (2004a) pour une discussion plus approfondie de v^m dans un contexte d'échantillonnage pur.

Pour un plan de sondage général, nous pouvons nous approcher d'un bon estimateur de la variance/erreur quadratique moyenne avec

$$v^{\text{com}} = \sum_{k \in S} (w_k^k - w_k^k) x_k^k + \sum_{k \in S} \sum_{j \in S, k \neq j} [(\pi_{kj}^k / \pi_k^k \pi_j^j) - \pi_k^k \pi_j^j] (w_k^k t_{kj}^k) (w_j^j t_{kj}^j). \quad (17)$$

Le deuxième membre de l'équation (17) estime le deuxième membre de l'équation (15) avec t_{kj}^k remplaçant e_k^k . Notons que $\sum_U (1 - \pi_k^*) e_k^k / \pi_k^*$ dans l'équation (15) est estimé par $\sum_S (w_k^k w_j^j - w_k^k) t_{kj}^k$ plutôt que par $\sum_S w_k^k (1 - \pi_k^*) t_{kj}^k$, ce qui rendrait v^{com} plus convergent avec v^{SSW} de l'équation (8). Cette substitution donne un estimateur de variance ayant de bonnes propriétés basées sur le modèle prédictif quand les e_k^k sont non corrélés, et $\sigma_k^k = \mathbf{x}_k^k \zeta$, pour un certain ζ . Elle peut être faite même en l'absence de non-réponse.

Lorsque l'échantillon réel comprend plusieurs degrés et peut être faite même en l'absence de non-réponse. Dans l'équation (11), le jackknife, v_j , peut être calculé à l'aide des poids de rééchantillonnage jackknife :

$$w_k^{(a_j)} = w_k^k a_k^{(a_j)} / a_k^k + \left(\sum_{m \in U} \mathbf{x}_m^m - \sum_{m \in S} w_m^m [a_m^{(a_j)} / a_m^k] \mathbf{x}_m^m \right) \times \left(\sum_{m \in S} a_m^{(a_j)} f(\mathbf{h}_m \mathbf{q}) \mathbf{h}_m^m \mathbf{x}_m^m \right)^{-1} a_k^{(a_j)} f(\mathbf{h}_k \mathbf{q}) \mathbf{h}_k^k \mathbf{x}_k^k. \quad (18)$$

ce qui est une généralisation évidente des poids de rééchantillonnage jackknife de l'équation (12). De nouveau, si $f(\theta) = 1$, v_j peut être calculé comme s'il n'y avait pas de non-réponse.

7. Modélisation de la couverture

Folsom et Singh (2000) ont fait remarquer que le traitement de la non-réponse au moyen de la pondération par calage peut aussi être appliqué pour corriger pour le sous-dénombrement. Dans le contexte, la phase quasi

quasi-randémisation au modèle prédictif, le même modèle de quasi-randémisation hypothétique s'applique à toutes les variables de l'enquête.

Des choix promoteurs pour $f(\cdot)$ sont $\exp(\cdot)$ et $1 + \exp(\cdot)$, ce dernier correspondant à un modèle de probabilité de réponse ajusté au moyen d'une fonction logistique de $\mathbf{h}^k \phi$. Il pourrait également être raisonnable de supposer que $h^k = x^{g_k}$ pour $\lambda < 1$. En particulier, fixer $\lambda = 0$ signifie que la probabilité que l'exploitation agricole k réponde à l'enquête annuelle sur les récoltes dépend uniquement du fait qu'elle ait déclaré du maïs, du blé ou des pommes de terre lors du recensement de l'agriculture précédent, plutôt que du volume déclaré de ces récoltes.

Dans l'exemple de l'enquête sur les récoltes, les composantes de \mathbf{x}^k provenant du recensement précédent étaient les meilleurs prédicteurs disponibles des valeurs correctes pondantes pour l'enquête annuelle avant l'échantillonnage. Le fait que l'entreprise agricole k réponde à l'enquête est toutefois davantage une fonction de la superficie courante consacrée à la culture du maïs, si tant est qu'il y en ait, que d'une approximation précédente de cette valeur. Par conséquent, il est tentant d'introduire les valeurs d'enquête dans \mathbf{h}^k , plutôt que les valeurs de recensement correspondantes. Comme nous allons voir, cette procédure pose un problème théorique.

Sachant une $f(\cdot)$, la méthode itérative décrite à la section 4 permettra souvent de découvrir un vecteur ligne \mathbf{q} tel que $T_x = \sum_{k \in S} a^k f(\mathbf{q}) \mathbf{x}^k$. Le cas échéant, l'estimation de T_y au moyen de $t_{y-CAL} = \sum_{k \in S} w_k y^k$, où $w_k = a^k f(\mathbf{h}^k \mathbf{b})$, aura de bonnes propriétés sous le modèle prédictif linéaire $y^k = \mathbf{x}^k \beta + \varepsilon_k$, où $E(\varepsilon_k | \{\mathbf{x}^g, \mathbf{h}^g, I^g | g \in U\}) = 0$ pour tout $k \in U$, $I^k = 1$ si l'élément k est présent dans l'échantillon original et qu'il répond, et 0 autrement.

L'absence de biais dans le modèle prédictif est simplement due au fait que les poids satisfont l'équation de calage. Notons toutefois que, si des composantes de \mathbf{h}^k proviennent de l'enquête plutôt que de \mathbf{x}^k , l'hypothèse du modèle prédictif voulant que $E(\varepsilon_k | \mathbf{h}^k) = 0$ peut être problématique. Dans les conditions extrêmes, considérons le cas où une telle composante est y^k proprement dit. Habituellement, les récoltes décrit plus haut, y^k peut être la superficie annuelle consacrée à la culture du maïs de l'exploitation agricole k . L'introduction de cette valeur dans \mathbf{h}^k rend biaisé l'estimateur par calage connexe pour le modèle prédictif pour le maïs.

Cependant, lorsque le modèle prédictif est correct (en traitant $E(\varepsilon_k | \{\mathbf{x}^g, \mathbf{h}^g, I^g | g \in U\}) = 0$ comme une partie intégrante du modèle), la pondération par calage fondée sur tout choix de $f(\cdot)$ produira des estimateurs ayant de bonnes propriétés fondées sur le modèle prédictif. Ces estimateurs auront aussi de bonnes propriétés fondées sur le modèle de

6.2 Estimation de l'erreur quadratique moyenne sous quasi-randémisation

Que l'on puisse déclarer raisonnablement ou non que t_{y-CAL} est sans biais par rapport au modèle prédictif n'a aucun effet sur ses propriétés sous quasi-randémisation. Notons que $\mathbf{h}^k \phi$ et $\mathbf{h}^k \mathbf{q}$ sont des valeurs scalaires et non des vecteurs. Puisque $T_x = \sum_{k \in S} a^k f(\mathbf{h}^k \mathbf{b}) \mathbf{x}^k$, nos hypothèses et le théorème de la valeur moyenne ($f(\mathbf{h}^k \mathbf{b}) = f(\mathbf{h}^k \phi) + f'(\mathbf{h}^k \phi) \mathbf{b} + f''(\mathbf{h}^k \phi) \mathbf{b} \mathbf{b}^T / 2 + \dots$) révèlent

$$T_x - \sum_{k \in S} a^k f(\mathbf{h}^k \phi) \mathbf{x}^k = \sum_{k \in S} a^k [f'(\mathbf{h}^k \phi) \mathbf{b} + f''(\mathbf{h}^k \phi) \mathbf{b} \mathbf{b}^T / 2 + \dots] \mathbf{x}^k = \mathbf{O}^p(N / \sqrt{n})$$

pour une grandeur scalaire θ^k comprise entre chaque $\mathbf{h}^k \mathbf{q}$ et $\mathbf{h}^k \phi$. Il découle de cela que, si $\sum_{k \in S} a^k f'(\mathbf{h}^k \phi) \mathbf{h}^k \mathbf{x}^k / N$ est inversible à la fois pour la population réalisée N et la limite de probabilité (rappelons que f est monotone, donc que f' n'est jamais nulle), alors

$$\mathbf{b} - \phi = \left\{ \sum_{j \in S} a^j f'(\mathbf{h}^j \mathbf{b}) \mathbf{h}^j \mathbf{x}^j \right\}^{-1} \left\{ T_x - \sum_{j \in S} a^j f'(\mathbf{h}^j \phi) \mathbf{h}^j \mathbf{x}^j \right\} = \mathbf{O}^p(1 / \sqrt{n}) + \mathbf{O}^p(1/n).$$

L'estimateur t_{y-CAL} a une erreur de

et

$$\cdot (\underline{u} \wedge N)^d \mathbf{O} = {}^w \partial {}^w \mathbf{q}^{(f \circ)w} v^S \Sigma$$

Par conséquent,

$$\begin{aligned} & \left(\frac{1}{\epsilon} u / N \right)^d O + \\ & \left(\gamma \partial^{\gamma_M} \sum^{-v_u} \gamma \partial^{\gamma_M} \sum^{(+v)} \right) ([I]^{-v_u} / v_u) = \\ & \gamma \partial^{\gamma_M} \sum^{-\gamma} \partial^{(f)\gamma_M} \sum^{-\gamma} = \text{TV} \circ \bar{\iota}_! - \text{TV} \circ \bar{\iota}_! \end{aligned}$$

et $\nu = \nu^{2LS} [I^d O + I]$ and $d \lim_{u \leftarrow \infty} (N / \nu^{2LS})$.

Les poids de rééchantillonnage définis par l'équation (12) ne nécessitent pas d'itération même si les poids de calage sont eux-mêmes produits de cette façon, ce qui est fort intéressant du point de vue informatique. Cela permet non seulement d'économiser du temps d'ordinateur, mais aussi d'éviter qu'une solution itérative puisse exister pour les w_i , mais non pour les poids de rééchantillonnage.

6.1 Modèle de quasi-randomisation et modèle prédictif

6. Non-réponse totale

A la présente section, nous examinons le traitement de la non-réponse totale (unité complète) en tant que phase supplémentaire de l'échantillonnage de Poisson. Il s'agit essentiellement d'un modèle de *quasi-randomisation*. Nous supposons que chaque élément k de l'échantillon original, maintenant dénoté F , a une probabilité de réponse, p_k . La probabilité que les éléments k et j répondent conjointement est $p_k p_j$, et le fait que l'élément k réponde (sachant un vecteur de covariables) est indépendant du fait qu'il soit choisi à partir de l'échantillon original.

Il est souvent possible de construire un ensemble de poids tel que l'estimateur par calage soit convergent par rapport au plan d'échantillonnage sous le modèle de quasi-randonisation. Nous recherchons ici un moyen particulier de construire ces poids. Pour cela, nous supposons que le modèle de quasi-randonisation est correct. Chaque élément est relié à un vecteur ligne de variables auxiliaires, \mathbf{x}_i , pour lequel $T_i = \sum_j v_j x_{ij}$ est connu. Enfin, nous supposons que chaque p_i est de la forme :

$$(13) \quad (\phi^{\gamma} \psi) f / I = {}^{\gamma} d$$

où ϕ est un vecteur colonne inconnu, \mathbf{h}_k est un vecteur ligne de même dimension que \mathbf{x}_k , et $\sum_{k=1}^N \mathbf{h}_k^T \mathbf{x}_k / N$, ou S représente maintenant le « sous-échantillon » de répondants, N , et pour la limite de probabilité.

connue, mais la valeur du paramètre qui la régit, ϕ , ne l'est pas. Lorsqu'elle est l'équation des poids de calage, $w_k = a^k f(\mathbf{h}^k \mathbf{b})$, de sorte que l'équation de calage proprement dite, $\sum_k w_k x_k = T_x$ soit vérifiée, $f(\mathbf{h}^k \mathbf{b})$ estime implicitement l'inverse de la probabilité de réponse de l'élément. Contrairement à la situation où le calage est utilisé pour la correction de l'écart de $\sum_k a^k x_k$ par rapport à T_x du purement à l'erreur d'échantillonnage, $f(0)$ et $f'(0)$ n'ont pas à valoir 1 et $\mathbf{h}^k \phi$ n'a pas à valoir zéro. Le choix le plus évident pour \mathbf{h}^k lorsqu'on possible le modèle de réponse donné par l'équation (13) est x_k proprement dit. Dans un exemple courant de pondération par calage pour corriger pour la non-réponse, les composantes de \mathbf{x}_k sont des variables indicatrices : $x_k^g = 1$ quand k est dans le groupe g et zéro sinon. Si les groupes sont mutuellement exclusifs, la pondération par calage équivaut à une pondération dans les classes de poststratification. Voir, par exemple, Särndal, Swensson et Wretman (1992, page 585). Le qualificatif « prédictif » est nécessaire pour distinguer ce modèle du modèle de quasi-randomisation) sous-jacent que chaque élément k du groupe g , qu'il répond ou non, a une moyenne courante : $E_g(Y_k) = \beta_g^g$. Le modèle de réponse quasi aléatoire est analogue : $P_k = 1/\phi^k$. Les deux modèles sont toutefois conceptuellement très différents.

Si les groupes ne sont pas mutuellement exclusifs, l'ajustement proportionnel itératif (API) est une méthode de détermination des poids de calage. Il en existe d'autres qui dépendent de la forme exacte de la fonction de réponse hypothétique (f). Le modèle prédicatif demeure linéaire, $E_g(Y_g) = \mathbf{x}_g'\beta$, tandis que le modèle de réponse qui donne lieu à l'API, $p_k = \exp\{-\mathbf{x}_k'\phi\}$, ne l'est pas. Berry, Flatt et Pierce (1996) donnent un exemple d'utilisation de l'API pour ajuster pour la non-réponse.

Dans de nombreuses applications de la pondération par calage, les composantes de \mathbf{x}_i sont continues ou semi-continues, plutôt que dichotomiques. Dans une enquête annuelle sur les récoltes, par exemple, soit x_{1i} la quantité de maïs récoltée lors du recensement de l'agriculture précédente par l'exploitation agricole i , la quantité de blé récoltée par cette exploitation, x_{2i} , la quantité de pommes de terre récoltées, et ainsi de suite. L'enquête annuelle sur les récoltes possède un modèle prédictif hypothétique pour la superficie consacrée à la culture du maïs par l'exploitation agricole i , y_{1i} , de la forme $y_{1i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_{1i}$. L'indice d'adéquation le maïs, β_1 , existe d'autres valeurs d'enquête d'intérêt, comme la superficie consacrée à la culture du blé, et éventuellement des modèles prédictifs hypothétiques pour

chacune.

Le modèle de réponse quasi aléatoire pour l'enquête sur les récoltes dépend des hypothèses émises au sujet de $f(\cdot)$ et de \mathbf{h}_k dans l'équation (13) avec \mathbf{h}_k éventuellement égal

Dans le cas d'un échantillon à plusieurs degrés, il est logique de permettre que, dans le modèle prédictif, ε_k et ε_j soient corrélés quand k et j sont des éléments de la même UPE, mais autrement pas. Si l'on peut ignorer la correction pour population finie, la variance fondée sur un modèle d'un

ou S^a denote l'échantillon de n^a unités dans la strate α ($\alpha = 1, \dots, A$), et U^α , la population de la strate contenant

$$\left(\frac{n_{u/f} m}{\sum_{S \in f} m} - \frac{\chi_{f,m}}{\chi_{f,m}} \right) \times$$

L'établissement de propriétés asymptotiques pour v_{SS}^{SW} sous échantillonnage aléatoire simple stratifié est un exercice simple. Dans le présent contexte, v_{SS}^{SW} se réduit à

Dans leur article, Deville et Samard ont, en réalité, examiné dans l'équation (3), avec \mathbf{u} remplaçant les x_j , la version différentielle est donnée dans Demandt et Rao (2004), où les a_j de l'équation sont remplacés par $a_j f'(\mathbf{u}, \mathbf{q}')$. Ces auteurs soulignent dans un commentaire accompagnant cette dernière expression que les trois versions des r_i sont asymptotiquement identiques jusqu'à $\mathbf{f}'(0) = \mathbf{f}'(0) = \mathbf{I}$ et $\mathbf{q}^* = \mathbf{q}^*$ est asymptotiquement égal à 0. Ces identités asymptotiques pourraient ne plus être vérifiées lorsque la pondération par calage est utilisée pour corriger pour la non-réponse, comme nous le verrons à la section

$$(6) \quad \mathbf{y}^i = \mathbf{y}^i - \mathbf{x}^i \left(\sum_{j \in \mathcal{S}} \mathbf{h}^f \mathbf{h}^f \mathbf{x}^j \right) \sum_{j \in \mathcal{S}} \mathbf{h}^f \mathbf{h}^f \mathbf{x}^j.$$

$$(6) \quad \cdot^f \mathbf{x}^f, \mathbf{q}^f v \sum_{I^-}^{S \ni f} \left(\cdot^f \mathbf{x}^f, \mathbf{q}^f v \sum_{I^-}^{S \ni f} \right)^{\chi_{\mathbf{x}} - \chi_{\mathcal{I}}} = \chi_{\mathcal{I}}$$

En utilisant des arguments parallèles à ceux de Deville et Samard (1992), v_{SSW}^{ssw} s'applique aussi de façon générale à t_{YCAL}^{YCAL} avec les poids de calage définis par l'équation (7) moyenne sous randomisation.

Le terme *plug-in* vient du fait que $\hat{\gamma}$ est « introduit dans » (plugged into) v^{ss} à la place des γ : $e = \gamma'x - \hat{\gamma}'x = \gamma'x - (\sum_{i=1}^n \hat{\gamma}_i' x_i) - \sum_{i=1}^n \hat{\gamma}_i' y_i$ pour l'estimation de l'erreur quadratique

$$(8) \quad \cdot ({}^f\mathcal{M})({}^k\mathcal{M})[{}^k\mathcal{U}/({}^f\mathcal{U} - \pi_{k_j}(\pi_k))] \sum_{j \in S} \sum_{k \in S} = \text{MSS}^{\mathcal{V}}$$

Samdal, Swensson et Wreiman (1989) ont proposé cet estimateur de la variance ou de l'erreur quadratique moyenne sous randomisation à modèle *plug-in* pour γ_{GREG} sous un plan d'échantillonnage arbitraire :

5. Estimation de la variance

estimer par calage $E_{\text{test}}^w[\sum_{k \in S} w_k^2 \epsilon_k^2]$ sous des contraintes faibles, où $S(i)$ est l'ensemble d'éléments échantillonnés dans l'UPe_i et S' est l'ensemble d'UPe sélectionnés au premier degré de l'échantillonnage.

$$(01) \quad \left\{ \frac{{}^v u}{\left(\frac{{}^y u {}^y m \sum_{{}^v s \ni y} {}^v s \sum_{{}^j s \ni y} {}^v s \right)}_2} \left({}^y u {}^y m \sum_{{}^v s \ni y} {}^v s \right) - \sum_{{}^v s \ni f} {}^v s \right\} \times$$

être élevée. Il n'est pas difficile de montrer que v_{ST}^2 est asymptotiquement indistinguable de l'estimateur de la variance par le jackknife :

$$(II) \quad \left\{ z^{(\text{TVC}^{-1}t - (f(n)\text{TVC}^{-1}t))} \sum_{f \in S}^n \right\} ({}^n u / [1 - {}^n u]) \sum_{\nu}^{[=n]} = f_A$$

où $t_{\gamma-\text{CAL}(\alpha_j)} = \sum_{k \in S} w_{k(\alpha_j)}^k \gamma^k$, et les poids de calage par rééchantillonnage jackknife sont

$$(71) \quad \epsilon^{\gamma} \mathbf{q}^{(f\alpha)\gamma} v \left(\underset{\Gamma-}{\overset{w}{\mathbf{x}}} \mathbf{q}^{(f\alpha)} v \overset{\S \equiv w}{\sum} \right) \times \\ \left(\overset{w}{\mathbf{x}} \underset{v}{\overset{(f\alpha)w}{v}} v \underset{\mathcal{M}}{\overset{\S \equiv w}{\sum}} \overset{(f\alpha)w}{\sum} \right) + \gamma v / \overset{(f\alpha)}{v} \gamma v \underset{\mathcal{M}}{\overset{(f\alpha)}{\sum}}$$

$$\sum_{k \in S} x_k = x \quad \text{contraintes de telle sorte que}$$

Soit $S(\alpha)$ l'ensemble d'*éléments* dans la strate α (et non les UPE comme S_α) et $S(\alpha_j)$ l'ensemble d'*éléments* dans l'UPE j de la strate α . Sous des contraintes faibles que

$$\begin{aligned} & \epsilon(N)^d \mathbf{O} = {}^w \mathbf{x}^y \mathbf{M}^{(fn)S} \mathbf{v}^S \mathbf{Z} \epsilon(u/N)^d \mathbf{O} = \\ & \left({}^n u / {}^y \mathbf{x}^y \mathbf{M}^{(\cdot n)S} \mathbf{Z} - {}^y \mathbf{x}^y \mathbf{M}^{(fn)S} \mathbf{Z} \right) \left([{}^n u] / {}^n u \right) = \\ & {}^w \mathbf{x} [{}^w \mathbf{v} / (fn)w \mathbf{v}] {}^w \mathbf{M}^S \mathbf{Z} - {}^w \mathbf{x} \mathbf{Z} \end{aligned}$$

3. Redéfinition des poids de calage

Dans leur définition originale des poids de calage, Deville et Särndal (1992) posaient comme condition que l'ensemble des poids de calage, $\{w_k | k \in S\}$, minimisent une certaine fonction de distance entre les membres de l'ensemble et les poids d'échantillonnage originaux, les a_k . Il est satisfaisant que l'équation de calage, $t_{y, CAL} = \sum_S w_k Y_k$, soit inversible, l'estimateur par calage, $t_{y, CAL} = \sum_S w_k Y_k$, était à la fois sans biais sous le modèle donné par l'équation (2) et habituellement convergent sous randomisation. Estévaou et Särndal (2002) ont proposé d'éliminer l'exigence que les poids de calage minimisent une fonction de distance. À la place, ils ont essentiellement proposé que les w_k soient seulement obligés de satisfaire l'équation de calage et d'avoir la « forme fonctionnelle » suivante :

(6) $w_k = a^k (1 + h^k(q))$

où h^k est un vecteur ligne de même dimension que x_k , tel que $\sum_S a_k h^k x_k$ est inversible, et q est un vecteur colonne de même dimension. L'équation (6) est une généralisation faible de (4), où h^k remplace effectivement $c^k x^k$. Il n'est pas difficile de voir que $q = [(\sum_S a_j h^j x_j)^{-1}]^T (T_x - \sum_S a_j x_j)$. En outre, sous des contraintes faibles que nous nous proposons vérifier, $t_{y, CAL} = \sum_S w_k Y_k = \sum_S a_j Y_j$ et $(T_x - \sum_S a_j x_j) = \sum_S a_j h^j x_j$ est convergent sous randomisation quand $t_{y, EXP}$ l'est. Il est sans biais sous le modèle prédictif linéaire donné par l'équation (2) quand $E(e_k | \{x_g, h_g | g \in S\}, \{t_g | g \in U\}) = 0$ pour tout $k \in U$. Cela suggère une définition de rechange des poids de calage : un ensemble de poids, $\{w_k | k \in S\}$, tel que

i. les w_k satisfassent l'équation de calage pour $\{x_k | k \in U\}$ et,

ii. $t_{y, CAL} = \sum_S w_k Y_k$ soit convergent sous randomisation quand $t_{y, EXP}$ l'est sous des contraintes faibles.

Cette définition est celle que nous utiliserons. Cette définition élargie de la pondération par calage s'avérera fort utile lors du calage pour la correction de la non-réponse ou des erreurs de couverture.

Il découle de notre nouvelle définition que le calage à forme fonctionnelle d'Estévaou et Särndal est, en réalité, une forme de pondération par calage. En nous inspirant de la théorie économétrique, nous donnons aux composantes de h^k qui ne sont pas des combinaisons linéaires des composantes de x^k le nom de « variables instrumentales ».

4. Calage éventuellement non linéaire

En partant des idées de Deville et Särndal (1992), nous pouvons généraliser la forme linéaire des poids de calage donnée par l'équation (6) à

(7) $w_{k, GEN} = a^k f(h^k(q^*))$

où f est une fonction monotone, dérivable deux fois avec $f(0) = 1$, $f'(0) = 1$ et $f''(0)$ est la dérivée première de f évaluée à 0) et q^* est choisi de sorte que l'équation de calage soit vérifiée. Contrairement à l'équation des poids de calage susmentionnée, l'équation de calage proprement dite, $\sum_S w_k x_k = T_x$, demeure linéaire. Notons que, puisque $f(0) = 1$, $f'(0) = 1$, $f(h^k(q^*)) \approx 1 + h^k(q^*)$. Strictement parlant, nous devrions utiliser un symbole supplémentaire sur $w_{k, GEN}$ (et plus tard sur $w_{k, LIN}$) pour dénoter le choix particulier de h^k . Nous l'avons laissé tomber pour simplifier.

Une solution, q^* , de l'équation (7) peut souvent être obtenue de façon itérative. On peut partir de $q^{(0)} = 0$, c'est-à-dire $\sum_S w_{k, GEN}^{(0)} = a^k f(0)$. Pour $r = 1, 2, \dots$, on fixe alors $q^{(r)} = b^{(r)} + \{[\sum_S f'(h^k(q^{(r-1)})) a^k h^k x_k]^{-1}]^T \times (T_x - \sum_S w_{k, LIN}^{(r-1)} x_k^r)\}$, et $w_{k, GEN}^{(r)} = a^k f(h^k(q^{(r)}))$. L'itération s'arrête à r^* quand $T_x = \sum_S w_{k, GEN}^{(r^*)} x_k^{r^*}$ à toutes fins utiles. Cependant, il faut se souvenir qu'il pourrait ne pas exister d'ensemble de poids pouvant être exprimé sous la forme de l'équation (7) tout en satisfaisant l'équation de calage.

Soulignons que $q^{(1)} = a^k (1 + h^k(q))$. Un développement en série de Taylor autour de zéro révèle $f(h^k(q^{(1)})) = 1 + h^k(q^{(1)}) + O_p(1/n)$ sous des contraintes faibles, de sorte que $\sum_S w_{k, LIN}^{(1)} Y_k = \sum_S w_{k, LIN} Y_k + O_p(N/n) = T_y [1 + O_p(1/n)]$. En outre, il n'est pas difficile de voir que $w_{k, GEN}^{(1)} = w_{k, LIN}^{(1)} [1 + O_p(1/n)]$, une égalité qui s'avère utile dans l'estimation de la variance.

L'exemple le plus courant en pratique d'une fonction non linéaire est $f(h^k(q)) = \exp(x_k(q))$, où les valeurs de chaque composante de x_k , dénotées x_{1k}, \dots, x_{pk} , sont 0 ou 1. Cela est effectivement la forme des poids de calage sur marges (APF) de Deming et Stephan (1940) calculés par ajustement proportionnel itératif. De nombreux auteurs ont constaté que la routine itérative décrite plus haut peut être utilisée même si les composantes de x^k ne sont pas binaires, comme elles le sont dans Deming et Stephan. Soulignons que les poids de calage par *raking généralisé* résultants sont systématiquement non négatifs.

2. Pondération par calage et l'estimateur GREG

Supposons que nous connaissions la probabilité de sélection, π_k , pour chaque élément d'échantillonnage k dans l'échantillon S . Nous pouvons estimer le total de la population, $T_y = \sum_U y_k$, où U dénote la population, au moyen de l'estimateur à facteur d'extension $t_{y,E} = \sum_S y_k / \pi_k = \sum_U y_k I_k / \pi_k$, où $I_k = 1$ quand $k \in S$ et 0 autrement. En traitant les I_k comme des variables aléatoires, il est facile de voir que $t_{y,E}$ est un estimateur sans biais de T_y . Les propriétés qui découlent du fait que les I_k sont traitées comme des variables aléatoires sont dites *fondées sur la randomisation*. Nous pouvons également écrire $t_{y,E} = \sum_U a_k y_k = \sum_S a_k y_k$, où $a_k = I_k / \pi_k$ est le poids d'échantillonnage de l'élément k .

Deville et Särndal (1992) ont inventé l'expression « estimateur par calage » pour décrire un estimateur de la forme $t_{y,CAL} = \sum_S w_k y_k$, où $\sum_S w_k x_k = \sum_U x_k = T_x$ pour un certain vecteur ligne de variables auxiliaires, $\mathbf{x}_k = (x_k^{(1)}, \dots, x_k^{(p)})$, au sujet duquel T_x est connu. Puisqu'il existe généralement un continuum d'ensembles $\{w_k | k \in S\}$ qui satisfont l'équation de calage :

$$(1) \quad \sum_{k \in S} w_k \mathbf{x}_k = T_x,$$

Deville et Särndal ont imposé comme condition que la différence entre l'ensemble de poids, $\{w_k | k \in S\}$, satisfaisant l'équation (1) et $\{a_k | k \in S\}$ minimise une fonction de perte.

Une autre approche de l'échantillonnage consiste à traiter les y_k comme des variables aléatoires satisfaisant le modèle prédicatif linéaire :

$$(2) \quad y_k = \mathbf{x}_k \beta + \varepsilon_k,$$

où $E(\varepsilon_k | \{\mathbf{x}_k, I_k | g \in U\}) = 0$ pour tout $k \in U$. En conditionnant cette espérance sur les I_k , nous supposons que l'on peut ignorer le mécanisme d'échantillonnage. Il s'agit d'un aspect critique, et parfois déraisonnable, du cadre (prédicatif) *fondé sur un modèle*.

Il est facile de voir que $t_{y,CAL}$ est un estimateur sans biais de T_y sous le modèle en ce sens que $E(t_{y,CAL} - T_y) = 0$ (en supprimant le conditionnement pour simplifier la notation); l'indice ε fait référence au traitement des ε_k comme des variables aléatoires (et des I_k comme des constantes prédéterminées).

Aux fins de notre étude, l'estimateur par la régression généralisée ou GREG a la forme :

$$t_{y,GREG} = t_{y,E} + \left(T_x - \sum_{k \in S} a_k \mathbf{x}_k \right) \left(\sum_{k \in S} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in S} a_k \mathbf{x}_k y_k, \quad (3)$$

où c_k est une constante arbitraire qui peut ou non être une fonction de \mathbf{x}_k , et $\lim_{N \rightarrow \infty} \sum_U c_k \mathbf{x}_k \mathbf{x}_k' / N = \mathbf{V}$ est une

$$w_k = a_k + \left(T_x - \sum_{j \in S} a_j \mathbf{x}_j' \right) \left(\sum_{j \in S} a_j \mathbf{x}_j \mathbf{x}_j' \right)^{-1} c_k a_k \mathbf{x}_k'.$$

Strictement parlant, les w_k sont des fonctions de l'échantillon réalisé, S , et des $c_k a_k$, mais nous supprimons cela dans la notation pour simplifier. Observons que les poids de calage peuvent être exprimés sous la forme

$$(4) \quad w_k = a_k(1 + c_k \mathbf{x}_k' \mathbf{q}),$$

où $\mathbf{q} = (\sum_S a_k c_k \mathbf{x}_k' / \mathbf{x}_j')^{-1} (T_x - \sum_S a_j \mathbf{x}_j')$ est un vecteur colonne, puisque $\mathbf{x}_k \mathbf{q}' = \mathbf{b} = \mathbf{b}' \mathbf{x}_k'$.

Supposons que des conditions de régularité raisonnables soient vérifiées (voir, par exemple, Kott 2004a pour un traitement plus approfondi) et que le plan d'échantillonnage est tel que $t_{y,E} - T_y = O_p(N/\sqrt{n})$, où n est la taille prévue de S (la taille réelle peut être aléatoire), $\sum_S a_k \mathbf{x}_k' = O_p(N/\sqrt{n})$, et $\sum_S a_k c_k \mathbf{x}_k' \mathbf{f}_k' = \sum_U c_k \mathbf{x}_k' \mathbf{f}_k' = O_p(N/\sqrt{n})$, où \mathbf{f}_k peut être \mathbf{x}_k ou y_k . Soit $\varepsilon_k = y_k - \mathbf{x}_k' (\sum_U c_k \mathbf{x}_k' \mathbf{x}_k')^{-1} \sum_U c_k \mathbf{x}_k' y_k$, de sorte que $\sum_U c_k \mathbf{x}_k' \varepsilon_k = 0$, et $\sum_S a_k c_k \mathbf{x}_k' \varepsilon_k = O_p(N/\sqrt{n})$. Nous pouvons exprimer l'erreur de $t_{y,GREG}$ sous la forme

$$\begin{aligned} t_{y,GREG} - T_y &= \sum_{k \in U} w_k y_k - \sum_{k \in U} y_k \\ &= \sum_{k \in S} w_k \varepsilon_k - \sum_{k \in U} \varepsilon_k \quad \text{car } \sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k \\ &= \sum_{k \in S} a_k \varepsilon_k + \left(T_x - \sum_{k \in S} a_k \mathbf{x}_k' \right) \left(\sum_{k \in S} a_k c_k \mathbf{x}_k' \mathbf{x}_k' \right)^{-1} \sum_{k \in S} a_k c_k \mathbf{x}_k' \varepsilon_k \\ &= \sum_{k \in U} a_k \varepsilon_k - \sum_{k \in U} \varepsilon_k + O_p(N/n). \end{aligned} \quad (5)$$

Il est maintenant aisé de voir que l'estimateur GREG est convergent sous randomisation, autrement dit, $p \lim_{n \rightarrow \infty} [(t_{y,GREG} - T_y) / N] = 0$. En outre, le biais relatif et l'erreur quadratique moyenne relative de l'estimateur GREG dans les conditions de randomisation sont d'ordre $1/n$. Puisque l'erreur quadratique moyenne = biais² + variance, nous pouvons conclure que le biais sous randomisation de l'estimateur GREG est habituellement un contributeur asymptotiquement non significatif à l'erreur quadratique moyenne de cet estimateur.

Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture

Philip S. Kott¹

Résumé

La pondération par calage peut être utilisée pour corriger la non-réponse totale et (ou) les erreurs de couverture sous des modèles appropriés de quasi-randomisation. Divers ajustements par calage qui sont asymptotiquement identiques dans un contexte d'échantillonnage pur peuvent diverger lorsqu'ils sont utilisés de cette manière. L'introduction de variables instrumentales dans la pondération par calage permet que la non-réponse (dissons) soit une fonction d'un ensemble de caractéristiques différentes de celles comprises dans le vecteur de calage. Si l'ajustement par calage a une forme non linéaire, une variante du jackknife permet d'éliminer le besoin d'itération dans l'estimation de la variance.

Mots clés : Modèle prédictif; modèle de quasi-randomisation; convergent sous quasi-randomisation; variable instrumentale; ajustement proportionnel itératif (raking) généralisé.

1. Introduction

La méthode de pondération par calage a été mise au point au départ en vue de réduire les erreurs d'échantillonnage en maintenant la convergence sous randomisation. Deville et Särndal (1992) ont démontré que de nombreuses formes de pondération par calage sont asymptotiquement identiques dans le contexte de l'échantillonnage, ce qui a fait progresser grandement notre compréhension des méthodes courantes de repondération, telles que la méthode itérative du quotient aussi appelée l'ajustement proportionnel itératif (APL, "raking" en anglais), qui ne se trouve pas sous le format de l'estimateur par la régression généralisée (GREG).

Folsom et Singh (2000) ont montré que la pondération par calage peut aussi être utilisée pour corriger les erreurs connues de couverture et (ou) la non-réponse totale sous des modèles appropriés de quasi-randomisation. Ces travaux n'ont été publiés dans aucune revue avec comité de lecture. Le cœur du présent article est une répétition des principaux résultats publiés dans Folsom et Singh, y compris une modification nécessaire de l'approche de Deville-Särndal en vue de modéliser l'estimation de la variance ou de l'erreur quadratique moyenne sous randomisation dans ce contexte élargi. Une version antérieure, strictement linéaire, de la pondération par calage pour l'ajustement pour la non-réponse totale peut être consultée dans Fuller, Loughlin et Baker (1994). Voir aussi Lundström et Särndal (1999). Nous faisons une distinction entre le modèle prédictif qui sous-tend habituellement le calage et le modèle de quasi-randomisation de Folsom et Singh. Toutefois, contrairement à ces deux auteurs, nous examinons ici les propriétés dans les deux cas. En outre, les variables explicatives du modèle de quasi-randomisation peuvent différer des variables de

calage, ce qui est également permis dans Lundström et Särndal.

Nous proposons un nouveau jackknife qui est analogue à l'estimateur de la variance par linéarisation de Deville-Särndal. Il repose sur l'utilisation de poids de rééchantillonnage calculés en une étape, quoique les poids de calage proprement dits puissent être déterminés itérativement.

Après la présentation de la notion bien connue de pondération par calage, à la section 2, nous passons en revue le cas particulier de l'estimateur GREG dans un contexte d'échantillonnage pur. À la section 3, nous décrivons l'extension de la pondération par calage d'Eustevao et Särndal (2000) dans sa forme linéaire, afin d'inclure des variables instrumentales. À la section 4, nous étendons le traitement de la pondération par calage de Deville et Särndal, afin d'inclure la possibilité de variables instrumentales. À la section 5, nous passons en revue l'estimation de la variance ou de l'erreur quadratique moyenne, et proposons un nouveau jackknife pour certains plans d'échantillonnage. À la section 6, nous décrivons comment la pondération par calage peut être utilisée pour la correction de la non-réponse. Dans ce contexte, les diverses formes fonctionnelles de la pondération par calage ne doivent plus nécessairement être asymptotiquement identiques. À la section 7, nous discutons des modèles de quasi-randomisation pour les erreurs de couverture, c'est-à-dire le sous-ou suréchantillonnage dans la base de sondage. À la section 8, nous donnons un petit exemple empirique appuyant le nouveau jackknife. Enfin, à la section 9, nous présentons une discussion des diverses approches et des domaines dans lesquels les travaux de recherche doivent se poursuivre.

Bibliographie

- Baker, M., McNicholas, A., Garrett, N., Jones, N., Stewart, J., Koberstein, V. et Lemmon, D. (2000). Household crowding: A major risk factor for epidemic meningococcal disease in Auckland children. *Pediatric Infectious Disease Journal*, 19, 983-990
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.
- Breslow, N.E. (1996). Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91, 14-28.
- Breslow, N.E. (2004). Case-control studies. Dans *Handbook of Epidemiology*. (Eds. W. Ahrens et I. Pigeot). New York : Springer, 287-319.
- Breslow, N.E., et Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, 75, 11-20.
- Breslow, N.E., et Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Applied Statistics*, 48, 457-468.
- Breslow, N.E., et Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase outcome-dependent sampling. *Journal of the Royal Statistical Society*, B, 59, 447-461.
- Brogan, D.J., Denniston, M.M., Litt, J.M., Flag, E.W., Coates, R.J. et Britton, L.A. (2001). Comparison of telephone sampling and area sampling: Response rates and within-household coverage. *American Journal of Epidemiology*, 153, 1119-1127.
- Cosslett, S.R. (1981). Maximum likelihood estimation for choice-based samples. *Econometrica*, 49, 1289-1316.
- DiGaetano, R., et Waksberg, J. (2002). Trade-offs in the development of a sample design for case-control studies. *American Journal of Epidemiology*, 155, 771-775.
- Fears, T.R., et Brown, C.C. (1986). Logistic regression models for retrospective case-control studies using complex sampling procedures. *Biometrics*, 42, 955-960.
- Fears, T.R., et Gail, M.H. (2000). Analysis of a two-stage case-control study with cluster sampling of controls: Application to non-melanoma skin cancer. *Biometrics*, 56, 190-198.
- Grabaud, B.L., Fears, T.R. et Gail, M.H. (1989). Effects of cluster sampling on epidemiologic analysis in population-based case-control sampling. *Biometrics*, 45, 1053-1071.
- Hartge, P., Britton, L.A., Rosenthal, J.F., Cahill, J.L., Hoover, R.N. et Waksberg, J. (1984). Random digit dialing in selecting a population-based control group. *American Journal of Epidemiology*, 120, 825-833.
- Hartge, P., Britton, L.A., Cahill, J.L., West, D., Hank, M., Austin, D., Silverman, D. et Hoover, R.N. (1984). Design and methods in a multi-center case-control interview study. *American Journal of Public Health*, 74, 52-56.
- Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys*. New York : John Wiley & Sons, Inc.
- Lawless, J.F., Kalbfleisch, J.D. et Wild, C.J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society*, B, 61, 413-38.
- Lee, A.J., Scott, A.J. et Wild, C.J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika*, 95 (A partire).
- Manski, C.F., et McFadden, D. (Eds) (1981). *Structural Analysis of Discrete Data with Econometric Applications*. New York : John Wiley & Sons, Inc.
- Miettinen, O.S. (1985). The case-control study: Valid selection of subjects. *American Journal of Epidemiology*, 135, 1042-1050.
- Prentice, R.L., et Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.
- Neuhäus, J., Scott, A.J. et Wild, C.J. (2002). The analysis of retrospective family studies. *Biometrika*, 89, 23-37.
- Neuhäus, J., Scott, A.J. et Wild, C.J. (2006). Family-specific approaches to the analysis of retrospective family data. *Biometrics*, 62, sous presse.
- Rao, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.
- Rao, J.N.K., Scott, A.J. et Skinner, C.J. (1998). Quasi-score tests with survey data. *Statistica Sinica*, 8, 1059-1070.
- Scott, A.J., et Wild, C.J. (1986). Fitting logistic models under case-control or choice-based sampling. *Journal of the Royal Statistical Society*, B, 48, 170-182.
- Scott, A.J., et Wild, C.J. (2001a). The analysis of clustered case-control studies. *Applied Statistics*, 50, 57-71.
- Scott, A.J., et Wild, C.J. (2001b). Fitting regression models to case-control data by maximum likelihood. *Journal of Statistical Planning and Inference*, 96, 3-27.
- Scott, A.J., et Wild, C.J. (2002). On the robustness of weighted likelihood. *Journal of the Royal Statistical Society*, B, 64, 207-220.
- Wacholder, S., McLaughlin, J.K., Silverman, D.T. et Mandel, J.S. (1991). Selection of controls in case-control studies. I Principles. *American Journal of Epidemiology*, 135, 1019-1028.
- Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.
- Waksberg, J. (1998). Random digit dialing sampling for case-control studies. Dans *Encyclopedia of Biostatistics*. (Eds. P. Armitage et T. Colton). New York : John Wiley & Sons, Inc., 3678-3682.
- Wienisch, M., Lee, M., Milke, R., Newman, B., Barger, G., Davis, R., Wiens, J. et Neuhäus, J. (1997). Familial and personal medical history of cancer and nervous system conditions among adults with glioma and controls. *American Journal of Epidemiology*, 145, 581-93.

nous pouvons identifier les familles contenant plus d'un cas, il serait alors possible d'atteindre une efficacité sensiblement plus grande en suréchantillonnant fortement ces familles. Essentiellement, nous considérons la famille comme l'unité d'échantillonnage; définissons une « famille-sélectionnons un échantillon cas-témoins de familles. Il s'agit d'un domaine important où de nombreux travaux restent à accomplir.

10. Conclusion

L'étude cas-témoins sur la population est l'un des domaines où la pratique a devancé la théorie. Avant que je sache, le seul ouvrage où le sujet est abordé en profondeur est celui de Korn et Graubard (1999, chapitre 9). Un aspect auquel a été accordée une attention théorique raisonnablement importante dans la littérature est la stratification. Des méthodes efficaces en vue d'intégrer des variables de stratification dans l'analyse ont été élaborées, entre autres, par Scott et Wild (1997), Breslow et Holubkov (1997), ainsi que Lawless et coll. (1999), dans des circonstances où les variables peuvent prendre uniquement un ensemble fini de valeurs. Breslow et Chattejee (1999) ont examiné le meilleur moyen d'utiliser ce genre d'information à l'étape de la conception de l'étude. L'extension de tous ces travaux (analyses ainsi que conception) à des situations où nous possédons de l'information sur des variables continues, comme l'âge, pour tous les membres de la population est un domaine où les travaux doivent se poursuivre. Bien que l'échantillonnage à plusieurs degrés soit d'usage répandu, l'effet de la mise en grappes a fait couler nettement moins d'encre. Fout exception Graubard et coll. (1989), Fears et Gail (2000), ainsi que Scott et Wild (2001a). Le présent article suscitera peut-être d'autres travaux portant sur ce sujet important. En particulier, puisque le problème se résume essentiellement à l'estimation de deux moyennes de population (voir l'équation (8)), il devrait être possible d'appliquer une grande partie des connaissances sur les plans de sondage efficaces à la résolution de ce problème.

Remerciements

Je tiens à remercier les examinateurs, ainsi que Barry Graubard et Graham Kalton, dont la discussion réfléchie d'une version antérieure du présent article a fait progresser considérablement ma compréhension du sujet. Enfin, j'aimerais remercier tout spécialement mes collaborateurs de longue date Chris Wild, avec lequel ont été réalisés presque tous les travaux qui sous-tendent le présent article, et Jon Rao, auquel je dois essentiellement tout mon savoir sur l'analyse des données d'enquête.

Dans Neuhaus et coll. (2006), nous élaborons des méthodes semi-paramétriques efficaces pour l'échantillonnage stratifié à plusieurs degrés dans des situations où la stratification dépend de la réponse, éventuellement d'une façon non spécifiée qui doit être modélisée, et les observations dans une unité primaire d'échantillonnage sont reliées au moyen d'un modèle paramétrique. Le calcul des estimations requiert la résolution des $p + 1$ équations d'estimation, où p est la dimension du vecteur de paramètres. La matrice de covariance peut aussi être estimée facilement en utilisant une analogie de l'inverse de la matrice d'information observée. La procédure complète peut être exécutée à l'aide d'une routine de maximisation raisonnablement générale, mais demande néanmoins une certaine expertise en calcul.

Nous pourrions aussi ajuster les mêmes modèles en utilisant des estimateurs pondérés par les poids de sondage, ce qui offre l'énorme avantage de ne nécessiter aucun logiciel spécialisé. Dans notre exemple, les familles comprenant un cas auraient un poids de 1 et les familles comprenant un témoin auraient un poids de $1\ 942\ 490/462 \approx 4\ 200$. Étant donné cette grande différence, nous pourrions nous attendre à ce que les estimations pondérées soient très inefficaces. Malheureusement, il s'avère presque impossible d'ajuster un modèle intéressant pour lequel les estimations pondérées convergent. L'un des problèmes est que les estimations pondérées sont fondées presque entièrement sur l'échantillon de témoins et que l'on possède fort peu d'information au sujet des effets familiaux dans les familles de témoins (un autre problème est que nous ne possédons pas d'information sur l'âge des membres de la famille et que toute spécification du modèle sans la variable d'âge était exagérément incorrecte). Donc, nous avons dû recourir à une simulation, qui est loin d'être achevée à ce stade. Il semble cependant qu'ici l'efficacité des estimations pondérées soit inférieure à 10 % des estimations par la méthode semi-paramétrique du maximum de vraisemblance. Plus de détails sont donnés dans Neuhaus et coll. (2002, 2006).

Bien que nos simulations en soient à un stade très précoce, il est possible de tirer quelques conclusions provisoires. La principale est que les grandeurs intrafamiliales sont fort mal estimées, même en utilisant des méthodes entièrement efficaces. Les plans d'étude familiale cas-témoins, où l'information sur les membres de la famille est obtenue à titre de supplément à un plan cas-témoins standard, ne fournissent tout simplement pas suffisamment d'information pour estimer les paramètres qui intéressent les épidémiologistes génétiques, à moins que les associations soient extrêmement (voire déraisonnablement) fortes (il convient de souligner que tous les épidémiologistes génétiques ne sont pas d'accord sur ce point). L'utilisation des variantes plus efficaces est néanmoins possible. Ainsi, si

9. Études familiales cas-témoins

Si nous nous intéressons principalement aux paramètres du modèle marginal (1), alors les méthodes dont nous avons discuté aux sections précédentes sont faciles à appliquer et raisonnablement efficaces. L'utilisation de méthodes entièrement efficaces requiert la construction de modèles paramétriques de la dépendance intragappe et l'effort supplémentaire que cela demande en vaut rarement la peine. Cependant, il existe des situations où la structure de dépendance présente un intérêt intrinsèque. En particulier, il est de plus en plus fréquent que les épidémiologistes génétiques étouffent les données d'une étude cas-témoins standard au moyen d'information sur les réponses et les covariables fournie par des membres de la famille, afin d'essayer d'obtenir des renseignements sur le rôle de la génétique et de l'environnement. Cette approche peut être considérée comme un échantillonnage en grappes stratifié, où les familles sont les grappes et, dans ce cas, la structure intragappe est de toute première importance. L'exemple qui suit est assez typique.

Exemple 3

Wrensch, Lee, Mittek, Newman, Barger, Davis, Wiencke et Neuhaus (1997) ont réalisé une étude cas-témoins sur la population du gliome, forme la plus fréquente de tumeur maligne du cerveau, dans la région de la baie de San Francisco. Ils ont recueilli des renseignements sur tous les cas de gliome diagnostiqués durant un intervalle de temps particulier et sur un échantillon comparable de témoins sélectionnés par la méthode de composition aléatoire. Ils ont également recueilli des renseignements sur la situation de tumeur du cerveau et sur les covariables auprès des membres de la famille des sujets sélectionnés dans l'échantillon cas-témoins original. L'étude portait sur 476 familles comptant un cas de tumeur du cerveau et 462 familles comptant un témoin.

Nous pourrions utiliser les méthodes dont nous venons de discuter pour ajuster un modèle marginal de la probabilité de devenir une victime du gliome, mais les chercheurs s'intéressaient avant tout à l'estimation des caractéristiques intrafamiliales. Une approche consisterait à ajuster un modèle logistique mixte comprenant un ou plusieurs effets familiaux aléatoires. Notons que, strictement parlant, le plan d'échantillonnage de l'exemple 3 n'est pas compris dans ce plan d'étude cas-témoins. Ici, la stratification est reliée à la variable réponse, mais n'est pas entièrement déterminée par cette dernière. La strate 1 contient les 476 familles dans lesquelles un cas a été diagnostiqué durant un petit intervalle de temps déterminé, tandis que la strate 2 contient les 1 942 490 autres familles, dont certaines comprennent des victimes du cancer du cerveau.

Dans les deux scénarios, la valeur de β_0 est fixée de sorte que la proportion de cas dans la population soit de 1 sur 400, c'est-à-dire $W_0 = 0,9975$. La densité globale de x est représentée à la partie supérieure du graphique et les densités conditionnelles pour les cas et les témoins le sont au bas du graphique. Les valeurs de x_i et B_{i1} sont données pour $\gamma = W_0$ (étiquetées « Population ») et $\gamma = 0,5$ (étiquetées « Biais »). La seconde valeur correspond à la pondération d'échantillon si nous tirons des nombres égaux de cas et de témoins. Manifestement, dans les deux scénarios, la pondération d'échantillon produit une estimation de la pente appropriée pour des valeurs de x situées plus à l'extrémité de la queue supérieure de la distribution (c'est-à-dire pour les personnes à haut risque) que dans le cas de la pondération égale.

Notons que, si nous sélectionnons des échantillons aléatoires simples de taille $n_0 = n_1 = 200$ à partir de la population de la figure 1 (a), l'efficacité relative de la pondération d'échantillon ne serait que d'environ 16 %, et le biais de petit échantillon serait de 0,24. Dans ce cas, même si nous prenions la valeur de population comme cible, la pondération par les poids de sondage produirait une erreur quadratique moyenne plus grande que la pondération d'échantillon.

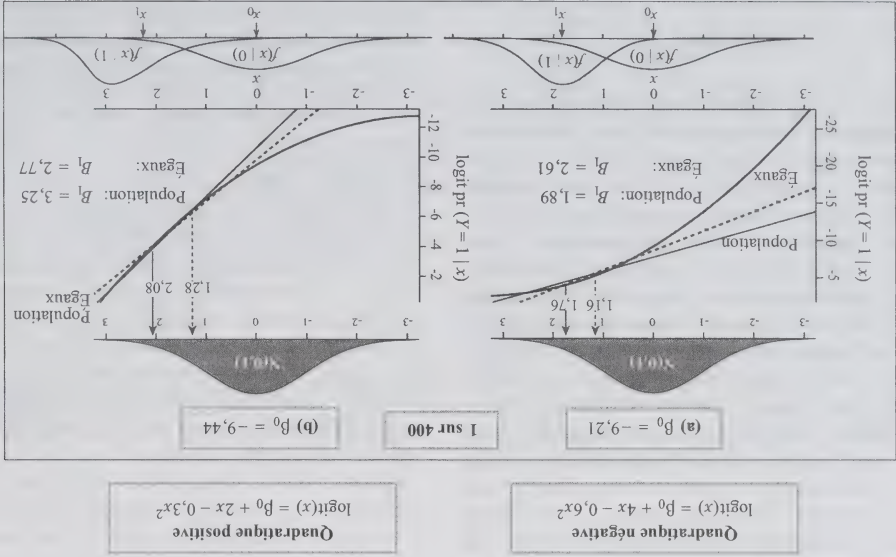
Un plus grand nombre de résultats sont présentés dans Scott et Wild (2002), où nous examinons aussi l'effet des covariables omises. Celles-ci s'avèrent avoir un effet semblable, mais un peu plus faible, que l'omission d'un terme quadratique. Quelle est la valeur de γ qu'il convient d'utiliser? Cela dépend clairement de l'utilisation que nous voulons faire du modèle résultant. Si notre principal intérêt est d'utiliser le modèle pour estimer des rapports de cotes à des valeurs de x où la probabilité d'un cas est élevée, et que l'échantillon est suffisamment grand pour que la variance et le biais de petit échantillon soient moins importants, nous pourrions utiliser les poids de population. Pour les tailles d'échantillon plus petites, ou si nous nous intéressons à des valeurs de x plus proches de la moyenne de population, les poids d'échantillon conviendraient mieux. Parfois, une valeur intermédiaire entre les poids de population et les poids d'échantillon pourrait représenter un compromis raisonnable. Par exemple, l'élagage des poids à 10 pour 1 (c'est-à-dire fixer $\gamma \approx 0,91$) dans l'exemple, au lieu de 1 pour 1 (pondération d'échantillon) ou 400 pour 1 (pondération de population) donne une efficacité de 70 % et un biais de petit échantillon de 0,04. Les valeurs correspondantes pour la pondération de population sont 16 % et 0,24. La valeur de $x_{0,91}$ est située presque exactement à mi-chemin entre $x_{0,5}$ et $x_{0,9975}$.

Par construction, l'estimateur pondéré en fonction de la population estime toujours l'approximation logistiqu linéaire que nous obtiendrions si nous disposions de données pour l'ensemble de la population. Par contre, ce que l'estimateur plus efficace pondéré en fonction de l'échantillon estime dépend des tailles d'échantillons particulières utilisées. Certaines personnes considéreraient cet élément à lui seul comme un raison suffisamment valable d'utiliser l'estimateur pondéré d'après la population et je soupçonne que la cible de leur inférence dépende du choix arbitraire de la taille d'échantillon.

Notre estimateur général β_0 , satisfaisant (10) converge vers la solution de l'équation (9), disons B_y , avec $\gamma = \lambda_0/(\lambda_0 + \lambda_1)$, qui dépend du modèle réel et de la distribution des covariables, ainsi que de γ . Dans Scott et Wild (2002), nous avons examiné ce qui arrive à B_y lors d'écarts faibles par rapport au modèle hypothétique (nous nous intéressons aux petits écarts, car en principe, les grands modèles qui devraient alors être améliorés en conséquence). Pour simplifier, supposons que nous ajustons un modèle linéaire ne contenant qu'une seule variable explicative pour le logarithme du rapport de cotes, mais que le modèle réel soit quadratique, disons

$$\logit\{P(Y = 1 | x)\} = \beta_0 + \beta_1 x + \delta x^2 \quad (19)$$

où δ est petit.



De toute évidence, la pente réelle de l'échelle logit, $\beta_1 + 2\delta x$, varie lorsque nous nous déplaçons le long de la courbe. Pour tout $0 < \gamma < 1$, B_y est égal à la pente réelle à un point donné sur la courbe. Denotons cette valeur par $x = x_y$. Soit x_0 la valeur attendue de x dans la population de témoins et soit x_1 la valeur attendue de x dans la population de cas. Nous supposons que $\beta_1 > 0$, de sorte que $x_0 < x_1$. Il s'avère que x_y est toujours compris entre x_0 et x_1 . Rappelons que la pondération par les poids de 0 à 1. Rappelons que la pondération par les poids de sondage correspond à $\gamma = W_0$ et que la pondération par les poids d'échantillon correspond à $\gamma = \omega_0 = n_0/n$. Habituellement, W_0 est sensiblement plus grand que ω_0 , de sorte que l'utilisation des poids de sondage donne une estimation de la pente pour des valeurs plus grandes de x_y , où la probabilité d'un cas est plus élevée, tandis que la pente estimée d'après la pondération d'échantillon s'approche davantage de la valeur moyenne de x dans la population. La figure 1, adaptée de Scott et Wild (2002), illustre la position dans deux scénarios, l'un avec une courbe positive et l'autre, une courbe négative, basés approximativement sur l'exemple 2. Nous choisissons une valeur de δ telle que celui-ci serait décalé à l'aide d'un test standard du rapport de vraisemblance dans environ 50 % des cas si nous sélectionnions des échantillons aléatoires simples de taille $n_0 = n_1 = 200$ à partir de la population.

La substitution de ces estimateurs aux moyennes d'échantillon dans l'équation (14) donne l'équation d'estimation

$$\hat{\gamma}_i(\beta) = \sum_{i=1}^h \sum_{j=1}^h w_{ij}^* \mathbf{x}_i (Y_i - P_{1h}(\mathbf{x}_i; \beta)) = \mathbf{0}, \quad (17)$$

avec $w_{ij}^* \propto \lambda_{ij} w_{ij} / \sum_{i=1}^h \sum_{j=1}^h w_{ij}$ pour les unités comprises dans $S_h^*(\ell = 0, 1; h = 1, \dots, H)$. Ce modèle peut être ajusté dans tout programme standard d'analyse de données

d'enquête en introduisant ces poids et l'information appropriée sur le plan de sondage. Notons que nous devons faire attention à la façon dont nous incluons ce que nous appelons des « strates » dans la spécification du plan. Si les unités primaires d'échantillonnage sont emboîtées dans les « strates », comme c'est le cas des régions géographiques dans l'exemple 1, il n'y a pas de problème et les strates doivent être incluses de la façon standard. Toutefois, si les unités primaires d'échantillonnage recourent les « strates », comme c'est le cas de l'âge dans l'exemple 1, et de l'âge et de l'ethnicité dans l'exemple 2, il ne s'agit plus de strates au sens habituel du terme en échantillonnage. Elles ne devraient pas être incluses dans les spécifications du plan,

mais simplement être traitées par la pondération. Parfois, nous voulons modéliser la contribution des variables de strate en utilisant une courbe paramétrique lisse au lieu de les inclure à l'aide de variables muettes. Par exemple, nous pourrions fort bien vouloir inclure une fonction linéaire de l'âge dans notre modèle, tant dans l'exemple 1 que dans l'exemple 2. La méthode de pondération par les poids de sondage et la pondération de comparaisons proposées à la section 6 s'appliquent l'une et l'autre, et aucun nouveau développement théorique n'est nécessaire. Par contre, les méthodes plus efficaces ne sont guères aussi simples. Des méthodes entièrement efficaces ont été élaborées dans la situation où des échantillons aléatoires simples de cas et de témoins sont tirés dans chaque strate (voir Scott et Wild 1997, ainsi que Breslow et Holubkov 1997), mais les équations d'estimation résultantes ne sont pas des combinaisons linéaires des moyennes de strates, et il n'existe aucune manière évidente de les généraliser à des plans d'échantillonnage plus complexes. Néanmoins, il existe un moyen un peu moins efficace, mais facile à étendre. Si nous modifions le modèle (14) en incluant $b_{\gamma_h} = \log(\lambda_{1h} W_{0h} / \lambda_{0h} W_{1h})$ comme correction, c'est-à-dire si nous supposons que

$$\log\{P^*(Y = 1 | \mathbf{x}, \text{Strate } h)\} = b_{\gamma_h} + \beta_{0h} + \mathbf{x}'\beta_1, \quad (18)$$

alors l'équation (15) produit des estimations convergentes, entièrement efficaces, de tous les coefficients, y compris $\beta_{0h} (h = 1, \dots, H)$. L'introduction des mêmes corrections dans les modèles ne contenant pas de terme β_{0h} et où le vecteur \mathbf{x} comprend des fonctions de la variable de stratification produit des estimateurs convergents pour tous

les coefficients avec une efficacité habituellement élevée (quoique non totale) (voir Fears et Brown 1986, ainsi que Breslow et Cain 1988). Cela se généralise à des plans de sondage arbitraires immédiatement. Il nous suffit d'utiliser l'équation (16) en remplaçant P_{1h} par P_{1h}^* défini en fixant $\logit(P_{1h}^*) = b_{\gamma_h} + \mathbf{x}'\beta_1$. Alors, tout programme d'analyse de données d'enquête qui permet d'appliquer des corrections peut être utilisé pour ajuster le modèle et fournir des estimations de types-erreurs, etc.

Quel est notre gain d'efficacité dans ce cas-ci? Nous avons exécuté plusieurs simulations, dont certaines sont décrites dans Scott et Wild (2002). La plupart des scénarios sont fondés sur l'étude de la méninigte de l'exemple 2 et nous fixons le ratio de la fraction d'échantillonnage de strate la plus grande à la plus faible dans l'échantillon de témoins à environ 10 pour 1. Sans aucune mise en grappes, le gain d'efficacité dû à l'utilisation de la méthode de correction (qui est le maximum de vraisemblance complète dans ce cas-ci) comparativement à la méthode *ad hoc* n'a jamais été supérieur à 10 %. Les efficacités relatives sont demeurées à peu près les mêmes lors de l'introduction d'une mise en grappes sur l'ensemble des strates. Quand nous sommes passés à la mise en grappes emboîtée dans les strates, les gains ont disparus progressivement à mesure que l'effet de plan augmentait et la méthode *ad hoc* est, en fait, devenue plus efficace que la méthode de correction, lorsque la valeur de l'effet de plan a atteint environ 1,5.

Comme nous l'avons mentionné plus haut, il est possible de produire des estimateurs semi-paramétriques entièrement efficaces si nous sommes prêts à modéliser la structure de dépendance à l'intérieur des unités primaires d'échantillonnage. Nous avons commencé à exécuter certaines simulations. Les premiers résultats donnent à penser que le travail supplémentaire que demande la modélisation ne vaudra presque jamais la peine si nous nous intéressons uniquement aux paramètres du modèle marginal (1). Notre conclusion provisoire est que les méthodes *ad hoc* partiellement pondérées (avec les poids d'échantillon) sont faciles à utiliser et donnent de suffisamment bons résultats pour la plupart des objectifs pratiques couverts par notre expérience, mais il s'agit toutefois d'un autre domaine où il conviendrait de poursuivre les travaux empiriques. Nous soulignerons cependant que, pour certains problèmes, comme l'étude familiale cas-témoins dont il est question à la section 9, le comportement intragrappe est intéressant en soi. Il faut alors recourir à des méthodes plus perfectionnées.

8. Robustesse

Il doit y avoir un piège quelque part. Que se passe-t-il si le modèle est incorrect? Quel est alors le prix du gain d'efficacité?

devons bien réfléchir à ce que nous essayons vraiment d'estimer. Nous examinons cette question à la section 8.

7. Échantillonnage stratifié

Le compromis proposé à la section précédente (c'est-à-dire utiliser la pondération standard par les poids de sondage dans les sous-populations définies d'après la situation de cas ou de témoin, mais combiner les sous-populations en utilisant les proportions d'échantillon) semble donner d'assez bons résultats en pratique, mais elle est entièrement efficace bien établies et faciles à appliquer. En particulier, si notre modèle comprend une ordonnée à l'origine distincte pour chaque strate, alors la régression logistique non pondérée ordinaire (avec un simple ajustement pour les ordonnées à l'origine de strate si l'on veut les obtenir) est la méthode semi-paramétrique efficace du maximum de vraisemblance (Prentice et Pyke 1979). Il est assez facile de l'étendre à des plans stratifiés plus généraux. Notre modèle est maintenant

$$\logit\{P(Y = 1 \mid \mathbf{x}, \text{Strate } h)\} = \beta_{0h} + \mathbf{x}^T \beta_{1h} \quad (14)$$

et l'équivalent stratifié de l'équation d'estimation (7) est

$$\sum_{h=1}^H \left(\lambda_{1h} \frac{\sum_{i \in \text{cas}} \mathbf{x}_i P_{0h}(\mathbf{x}_i; \beta) m_{1h}}{\sum_{i \in \text{témoins}} \mathbf{x}_i P_{1h}(\mathbf{x}_i; \beta) m_{0h}} - \lambda_{0h} \right) = 0. \quad (15)$$

À mesure que $m_{0h}, m_{1h} \rightarrow \infty$, la solution de (7) converge presque certainement vers la solution de

$$\sum_{h=1}^H (\lambda_{1h} E_{1h} \{ \mathbf{x} P_{0h}(\mathbf{x}; \beta) \} - \lambda_{0h} E_{0h} \{ \mathbf{x} P_{1h}(\mathbf{x}; \beta) \}) = 0, \quad (16)$$

avec l'extension évidente de la notation utilisée pour le cas non stratifié. Si le modèle (13) est vérifié, alors l'équation (8) a pour solution $\beta_{1h}^* = \beta_{1h}$ et $\beta_{0h}^* = \beta_{0h} + b_{jh}$ avec $b_{jh} = \log(\lambda_{1h} W_{0h} / \lambda_{0h} W_{1h})$. Puisque l'équation (14) ne fait intervenir que des moyennes de strate, nous pouvons estimer facilement ces dernières en utilisant les données provenant de tout plan de sondage raisonnable, par exemple

$$\hat{\mu}_h(\beta) = \frac{\sum_{i \in \text{cas}} \lambda_{1h} \mathbf{x}_{ih} (\lambda_{1h} - P_1(\mathbf{x}_{ih}; \beta))}{\sum_{i \in \text{cas}} \lambda_{1h} \mathbf{x}_{ih} (\lambda_{1h} - P_1(\mathbf{x}_{ih}; \beta))}.$$

par

Il nous reste à décider des bonnes valeurs de λ_1 et λ_0 . Nous pouvons souvent réaliser des gains importants en utilisant les poids d'échantillon ($\lambda_i = n_i/n$) plutôt que les poids de population ($\lambda_i = W_i$). Scott et Wild (2002) ont fait état de gains d'efficacité de 50 % ou plus dans l'exemple 2 et dans des simulations basées sur cette population. Les gains devenaient plus importants à mesure que s'intensifiait la force de la relation et qu'augmentait l'effet de la mise en grappe. De surcroît, la couverture des intervalles de confiance était plus proche de la valeur nominale en cas de pondération par les poids d'échantillon dans les simulations.

L'utilisation des poids d'échantillon est la stratégie la plus efficace disponible lorsqu'on a affaire à des échantillons aléatoires simples de cas et de témoins, mais pour des plans de sondage plus complexes, elle n'est plus entièrement efficace. Nous pourrions nous attendre à ce que des poids basés sur une forme de tailles d'échantillon équivalentes donnent de meilleurs résultats. Cette approche produit effectivement certains gains d'efficacité dans des simulations limitées décrites dans Scott et Wild (2001a). Toutefois, les gains sont relativement faibles, du moins lorsque l'effet de plan de l'échantillon de témoins est inférieur à 2, puisque $\text{Cov}\{\beta_{\lambda_i}\}$ est une fonction de λ_i très plate près de son minimum. Les considérations relatives à la robustesse dont nous discutons à la section 8 pourraient jouer un rôle plus important dans le choix de λ_i .

Les avantages qu'offre la pondération d'échantillon peuvent dépendre en grande partie du problème à l'étude. Korn et Graubard (1999, page 327) font remarquer que, dans leur expérience, la stratégie de pondération par les poids d'échantillon produit rarement de gros gains d'efficacité. De toute évidence, la poursuite des travaux, tant empiriques que théoriques, est nécessaire ici. Quoiqu'il en soit, il semble prudent d'ajuster le modèle en utilisant systématiquement les poids d'échantillon ainsi que les poids de population. Si les estimations des coefficients sont semblables, alors nous pouvons porter un jugement en nous basant sur les erreurs-types estimées. Par contre, des écarts significatifs entre les estimations des coefficients indiquent que le modèle a été mal spécifié. Si nous sommes incapables de corriger les déficiences du modèle, alors nous

d'échantillonnage sont sélectionnées indépendamment dans les strates (hypothèse qui est de toute façon celle faite dans tous les logiciels offrant l'approche de pondération par les poids de sondage), mais cela nous oblige à construire des modèles multivariés pour le vecteur des réponses dans une unité primaire d'échantillonnage. Le lecteur trouvera des renseignements détaillés dans Neuhäus, Scott et Wild (2002, 2006). À moins que nous ne nous intéressions à la structure intragappe en soi (comme dans les études familiales cas-témoins considérées à la section 9, par exemple), l'exercice demande beaucoup trop d'efforts pour être faisable, du moins dans les analyses de routine.

Pouvons-nous faire quelque chose de plus simple sans perdre trop d'efficacité? Naturellement, nous pouvons toujours nous rabattre sur les estimations pondérées. Cependant, elles sont tout aussi inefficaces pour les plans de sondage complexes que pour le cas simple examiné à la section précédente. En fait, nous pouvons obtenir de sensiblement meilleurs résultats sans trop de complications supplémentaires.

Revenons un instant à la situation de la section précédente où nous avons un échantillon aléatoire simple de taille n_1 provenant de la strate de cas et un échantillon aléatoire simple indépendant de taille n_0 provenant de la strate de témoins. Ici, toutes les unités de la strate ℓ ont un poids $w_i \propto W_i/n_i$, où W_i dénote la proportion de la population dans la strate, pour $\ell = 0, 1$. Si nous divisons tout au long par N et posons que $p_0(x; \beta) = 1 - p_1(x; \beta)$, alors nous pouvons réécrire l'équation (3) pour l'estimateur pondéré sous la forme

$$\hat{W}_1 = \frac{\sum_{\text{cas}} x_i p_0(x_i; \beta)}{\sum_{\text{cas}} x_i p_1(x_i; \beta)} - \frac{n_1}{W_0} \quad (6)$$

De même, nous pouvons écrire l'équation (5) pour l'estimateur efficace du maximum de vraisemblance sous la forme

$$\hat{\omega}_1 = \frac{\sum_{\text{cas}} x_i p_0(x_i; \beta)}{\sum_{\text{cas}} x_i p_1(x_i; \beta)} - \frac{n_1}{\omega_0} \quad (7)$$

où $\omega_i = n_i/(n_0 + n_1)$, pour $i = 0, 1$. Ces deux expressions sont des cas particuliers de l'ensemble général d'équations d'estimation

$$\lambda_1 = \frac{\sum_{\text{cas}} x_i p_0(x_i; \beta)}{\sum_{\text{cas}} x_i p_1(x_i; \beta)} - \frac{n_1}{\lambda_0} \quad (8)$$

À mesure que $n_0, n_1 \rightarrow \infty$, sous des contraintes faibles quant à la façon dont la population finie est générée à partir de la superpopulation, la solution de (8) converge presque certainement vers la solution β^* de

$$\lambda_1 E_1 \{X p_0(X; \beta^*) - \lambda_0 E_0 \{X p_1(X; \beta^*)\}\} = 0, \quad (9)$$

où $E_\ell\{\cdot\}$ dénote l'espérance conditionnelle sachant que $X = \ell$ pour $\ell = 0, 1$. Si le modèle (1) est vérifié, alors l'équation (8) a pour solution $\beta_1^* = \beta_1$ et $\beta_0^* = \beta_0 + b_1$ avec $b_1 = \log(\lambda_1 W_0 / \lambda_0 W_1)$ pour toute valeur positive de λ_0, λ_1 [voir Scott et Wild (1986) pour des détails de la preuve]. Donc, la solution de l'équation (8) produit des estimateurs convergents pour tous les coefficients de régression, sauf le terme constant pour tout $\lambda_\ell > 0$ ($\ell = 0, 1$). Comme dans le cas simple, il est facile de corriger les interférences au sujet du terme constant, à condition de connaître la proportion de cas dans la population. Maintenant, passons à des plans d'échantillonnage plus complexes. Puisque le premier membre de l'équation (9) ne fait intervenir que deux moyennes de sous-population, nous pouvons encore estimer ces moyennes pour tout plan de sondage standard. Cela suggère un estimateur, disons $\hat{\beta}_\lambda$, pour les plans d'échantillonnage généraux qui satisfait

$$\hat{S}_\lambda(\beta) = \lambda_1 \hat{t}_1(\beta) - \lambda_0 \hat{t}_0(\beta) = 0, \quad (10)$$

où $\hat{t}_\ell(\beta)$ est l'estimateur sur échantillon de la moyenne de sous-population $E_\ell\{X(1 - p - p^*)\}$ ($\ell = 0, 1$). La matrice de covariance de $\hat{\beta}_\lambda$ peut alors être obtenue à l'aide d'arguments standard de linéarisation, ce qui nous mène à une matrice de covariance estimée (« sandwich »)

$$\text{Cov}\{\hat{\beta}_\lambda\} \approx J_\lambda(\beta)^{-1} \text{Cov}\{\hat{S}_\lambda(\beta)\} J_\lambda(\beta)^{-1}, \quad (11)$$

avec $J_\lambda(\beta) = (-\partial \hat{S}_\lambda(\beta) / \partial \beta^T)$ et $\text{Cov}\{\hat{S}_\lambda(\beta)\} = \lambda_1^2 \text{Cov}\{\hat{t}_1(\beta)\} + \lambda_0^2 \text{Cov}\{\hat{t}_0(\beta)\}$. Ici, $\text{Cov}\{\hat{t}_\ell(\beta)\}$ dénote l'estimation par sondage habituelle qui devrait être disponible systématiquement pour tout plan de sondage standard, puisque $\hat{t}_\ell(\beta)$ est simplement une moyenne estimée.

Tout cela peut aussi être exécuté facilement au moyen de logiciels qui portent quel logiciel capable de traiter la régression logistique sous plan de sondage complexe, simplement en spécifiant le vecteur approprié de poids. Plus précisément, nous pouvons

$$\hat{t}_\ell(\beta) = \frac{\sum_{i \in S_\ell} w_i x_i (Y_i - p - p^*)}{\sum_{i \in S_\ell} w_i}, \quad (12)$$

où S_ℓ dénote la sous-population de cas (c'est-à-dire l'ensemble des unités pour lesquelles $X = 1$) et S_0 dénote la sous-population de témoins (l'ensemble des unités pour lesquelles $X = 0$). Alors, l'équation d'estimation (9) peut s'écrire sous la forme

$$\hat{S}_\lambda(\beta) = \sum_{i \in S_1} w_i^* x_i (Y_i - p_1(x_i; \beta)) = 0, \quad (13)$$

Les résultats s'étendent directement à l'échantillonnage aléatoire stratifié, à condition d'inclure dans le modèle une ordonnée à l'origine distincte pour chaque strate. De nouveau, il est possible d'obtenir des estimateurs semi-paramétriques efficaces de tous les coefficients, sauf les ordonnées à l'origine de strate en traitant simplement les données à l'aide d'un programme de régression logistique (non pondérée) ordinaire. De nouveau, les erreurs-types estimées et les inférences connexes sont également valides. Comme dans le cas de l'échantillonnage aléatoire simple, nous pouvons corriger les résultats pour les ordonnées à l'origine de strate à condition que nous connaissions les fractions d'échantillonnage dans les strates, mais, encore une fois, celles-ci ne présentent habituellement que peu d'intérêt.

Donc, dans ces situations simples, les estimations du maximum de vraisemblance sont plus faciles à calculer que les estimations pondérées, et elles sont aussi plus efficaces. Dans quelle mesure le sont-elles? Cela dépend du nombre de covariables, de la grandeur de leur coefficient et du ratio des fractions d'échantillonnage, mais la différence est souvent importante (ainsi, les estimations pondérées ont une efficacité d'environ 50 % dans l'exemple 2 de l'introduction, et de moins de 20 % dans l'exemple du cancer du cerveau que nous examinons à la section 8. Lawless, Kalbfleisch et Wild (1999) discutent de situations où l'efficacité est encore plus faible).

Enfin, nous notons que les estimations du maximum de vraisemblance présentent encore un autre avantage par rapport aux estimations pondérées : elles ont tendance à avoir de nettement meilleures propriétés sur petit échantillon, surtout quand l'efficacité des estimations pondérées est faible. Essentiellement, la pondération réduit la taille effective d'échantillon et cette dernière est l'élément qui détermine le moment où la théorie asymptotique commence à donner une bonne approximation (voir Scott et Wild 2001a pour plus de précisions). Maniètement, le prix de l'adhésion stricte aux poids de population peut être très lourd.

6. Échantillonnage plus complexe

Dans les deux exemples de la section 2, les témoins ont été obtenus par échantillonnage complexe à plusieurs degrés plutôt que par échantillonnage aléatoire simple. Comme nous l'avons mentionné dans l'introduction, cette situation est de plus en plus fréquente dans les études cas-témoins à grande échelle (à l'occasion, comme dans l'exemple 1, les cas sont également sélectionnés selon un plan d'échantillonnage complexe). Il est possible de dériver des estimateurs semi-paramétriques efficaces pour l'échantillonnage stratifié à plusieurs degrés, en supposant que les unités primaires

L'inconvénient de la pondération est, en général, qu'elle a tendance à être inefficace si les poids sont très variables (une règle empirique parfois proposée est que w_{\max}/w_{\min} ne devrait pas être supérieur à 10). Dans les études cas-témoins, la variation des poids est à peu près aussi extrême qu'elle peut l'être. Par exemple, le ratio de w_{\max} par rapport à w_{\min} est approximativement de 300 pour 1 dans l'exemple 1, et de 1 000 pour 1 dans l'exemple 2. Et des ratios encore plus extrêmes ne sont pas inhabituels. Aucun spécialiste chevronné de l'échantillonnage ne s'étonnerait de constater que la pondération n'est guère efficace dans ces circonstances.

Pouvons-nous trouver une solution plus efficace? La réponse est assurément affirmative dans certains cas. Des méthodes de vraisemblance entièrement efficaces ont été élaborées dans des situations où les cas et les témoins sont sélectionnés par échantillonnage aléatoire simple ou stratifié, et ces méthodes peuvent être nettement plus efficaces que les méthodes de pondération. Nous passons ces résultats en revue à la section suivante.

5. Révision : Cas simple

Examinons d'abord la situation la plus simple où les cas et les témoins sont sélectionnés par échantillonnage aléatoire simple et où nous ne disposons d'information au niveau de la population sur aucune des covariables à l'étape de l'élaboration du plan de sondage. Pour ce cas, il existe des méthodes semi-paramétriques entièrement efficaces d'estimation du maximum de vraisemblance bien établies. En outre, ces méthodes sont très faciles à appliquer en utilisant les logiciels standard (Prentice et Pyke 1979) (les méthodes sont semi-paramétriques, parce que la vraisemblance complète dépend de la distribution inconnue des covariables, ce que nous voulons en général modéliser).

En fait, il nous suffit d'ajuster le modèle (1) en utilisant un programme standard de régression logistique sans aucune pondération. Plus précisément, la résolution de l'équation non pondérée

$$\sum x_i' (y_i - p(x_i; \beta)) = 0, \quad (5)$$

échantillon

produit des estimations efficaces de tous les coefficients, sauf l'ordonnée à l'origine. Fait peut-être plus important, toutes les erreurs-types et les inférences résultantes que nous obtenons au moyen du programme standard sont également valides, de nouveau à l'exception de tout résultat où intervient l'ordonnée à l'origine. Il est assez simple de corriger les inférences dans lesquelles intervient l'ordonnée à l'origine, à condition que nous connaissions les fractions d'échantillonnage, mais nous ne nous intéressons souvent qu'aux autres coefficients de toute façon.

L'hypothèse que cette population finie est tirée à partir d'une superpopulation dans laquelle le modèle logistique conditionnel (1) est vérifié, alors nous pourrions estimer β en résolvant les équations d'estimation relatives à l'ensemble de la population, c'est-à-dire sous recensement, suivantes

$$(2) \quad S(\beta) = \sum_{i=1}^N x_i (y_i - p_i(x_i; \beta)) = 0,$$

où $p_i(x_i; \beta) = \frac{e^{\beta_0 + x_i^T \beta}}{1 + e^{\beta_0 + x_i^T \beta}}$. (II) s'agit des équations de vraisemblance si l'on suppose que les unités de population sont échantillonnées indépendamment à partir d'une superpopulation, mais que les estimateurs résultants sont convergents sous des structures de population nettement plus réalistes, à condition que le modèle (1) soit vérifié marginalement – voir Rao, Scott et Skinner 1998 pour une discussion plus approfondie.)

Maintenant, pour toute valeur fixée de β , $S(\beta)$ dans l'équation (2) est simplement un vecteur de population. Nous pouvons donc l'estimer d'après l'échantillon, disons par

$$(3) \quad \hat{S}(\beta) = \sum_{i=1}^{n_{\text{échantillon}}} w_i x_i (y_i - p_i(x_i; \beta)),$$

où w_i est l'inverse de la probabilité de sélection, peut-être corrigée de la non-réponse et de la poststratification. Fixer $\hat{S}(\beta)$ égal à 0 nous donne notre estimateur, $\hat{\beta}$. Nous pourrions appliquer la linéarisation ou le jackknife directement à $\hat{\beta}$ pour obtenir les erreurs-types. Nous pouvons aussi développer $\hat{S}(\hat{\beta})$ autour de la valeur réelle, β , et obtenir comme matrice de covariance estimée l'estimateur « sandwich »

$$(4) \quad \text{Cov}\{\hat{\beta}\} \approx \mathbf{J}(\hat{\beta})^{-1} \text{Cov}\{\hat{S}(\hat{\beta})\} \mathbf{J}(\hat{\beta})^{-1},$$

où $\mathbf{J}(\hat{\beta}) = -\partial S / \partial \beta^T = \sum_{i=1}^{n_{\text{échantillon}}} w_i p_i(x_i; \beta) p_0(x_i; \beta) x_i x_i^T$ avec $p_0 = 1 - p_i$. Puisque $S(\hat{\beta})$ est un vecteur de totaux, $\text{Cov}\{\hat{S}(\hat{\beta})\}$ devrait être disponible d'office pour tout plan de sondage standard. De nos jours, la plupart des grands logiciels statistiques (par exemple, SAS (PROC SURVEYLOGISTIC), SPSS (CSLOGISTIC), STATA (SVY:LOGIT), SUDAAN (LOGISTIC)) peuvent traiter de façon courante la régression logistique en cas d'échantillonnage complexe et de pondération. Par conséquent, il est assez simple de produire des estimations pondérées et de faire les inférences connexes.

Strictement parlant, dans notre cadre fondé sur un modèle, les probabilités de sélection sont souvent elles-mêmes des variables aléatoires basées sur une population finie hypothétiquement générée à partir du modèle. Nous pouvons tenir compte de ce fait en utilisant les résultats de Rao (1973), mais la correction est d'ordre $1/N$ et peut être ignorée dans la plupart des études de grande taille.

proportion de cas est faible, la différence est négligeable. Sinon, il est simple d'adapter les résultats qui suivent à cette variante – pour un développement rigoureux, voir Lee, Scott et Wild 2006). Habituellement, nous voulons ensuite utiliser les données d'échantillon pour ajuster un modèle de régression binaire de la probabilité marginale qu'une unité soit un cas ayant la forme d'une fonction des covariables. Le modèle utilisé est presque toujours logistique avec

$$(1) \quad \logit\{P(Y = 1 | \mathbf{x})\} = \log\left(\frac{P(Y = 1 | \mathbf{x})}{P(Y = 0 | \mathbf{x})}\right) = \beta_0 + \mathbf{x}^T \beta_1$$

disons, où β_0 et β_1 sont des paramètres inconnus et, tout au long de l'article, nous supposons que nous avons affaire au modèle (1). Les extensions à des modèles de régression plus généraux sont simples en principe (voir Scott et Wild 2001b), mais les expressions résultantes sont un peu moins élégantes que le modèle logistique.

Comment devons-nous y prendre pour ajuster le modèle (1) sachant les données d'échantillon? Les méthodes efficaces sont faciles à appliquer en cas d'échantillonnage aléatoire simple ou stratifié, mais nous nous intéressons ici à des méthodes d'échantillonnage plus complexes. Très souvent, on omet tout bonnement de tenir compte de l'échantillonnage complexe, ce qui risque d'entraîner tous les problèmes qui se posent habituellement lorsqu'on ne prend pas en compte la structure du plan d'échantillonnage. Des probabilités de sélection variables pourraient fausser la structure de la moyenne de sorte que les estimations produites par les programmes standard risquent d'être non cohérentes. La corrélation intragrappe pourrait réduire la taille effective d'échantillon, de sorte que les erreurs-types produites couramment seraient trop faibles, les intervalles de confiance, trop courts, les valeurs p , trop faibles, et ainsi de suite. Une stratégie simple qu'adoptent certains chercheurs pour réduire au minimum les effets consiste à garder petit le nombre de sujets dans chaque grappe (voir Granard, Fears et Gail 1989, par exemple). Cela réduit l'effet de plan, donc l'incidence sur la mise en grappes, mais le remède peut être coûteux. Dans les sections qui suivent, nous examinons certains moyens éventuels d'utiliser les plans d'échantillonnage standard, plus rentables.

4. Approche de la pondération de sondage

Une option évidente consiste à suivre l'approche standard des équations d'estimations pondérées qui est intégrée dans la plupart des logiciels contemporains d'analyse de données d'enquête (voir Binder 1983). Supposons d'abord que nous ayons des données provenant de la population finie complète. Si nous émettons

comparable de témoins ont été sélectionné parmi les autres enfants faisant partie de la population étudiée selon un plan d'échantillonnage complexe à plusieurs degrés. Au premier degré d'échantillonnage, 300 îlots de recensement (contenant chacun environ 70 ménages) ont été sélectionnés avec probabilité proportionnelle au nombre de maisons dans l'îlot. Au deuxième degré, un échantillon systématique de 20 ménages a été tiré dans chaque îlot sélectionné et les enfants provenant de ces ménages ont été choisis pour l'étude avec une probabilité variant selon l'âge et l'ethnicité, déterminée de façon qu'elle corresponde à la fréquence prévue parmi les cas. Les probabilités de sélection sont présentées plus loin au tableau 1 (l' signifie origininaire des îles du Pacifique). La taille des échantillons de grappes varie de 1 à 6, et environ 250 témoins ont été sélectionnés en tout. Cela correspond à une fraction d'échantillonnage d'environ 1 pour 400, en moyenne, de sorte que les cas ont été échantillonnés à un taux d'environ 400 fois celui des témoins.

Ces deux études sont assez représentatives de celles dont nous voulons discuter. Elles illustrent aussi les deux principales méthodes d'échantillonnage utilisées, à savoir la composition aléatoire et l'échantillonnage aréolaire. Une vive discussion des mérites relatifs de ces deux stratégies figure dans Broman, Demiston, Liff, Flagg, Coates et Brinton (2001), ainsi que dans DiGiacano et Waksberg (2002).

3. Conditions générales

Tableau 1		Probabilités de sélection	
ÂGE	MAORI ÎLES DU PACIFIQUE AUTRE		
≤ 1 an	0,29	0,70	0,10
≤ 3 ans	0,15	0,50	0,07
≤ 5 ans	0,15	0,31	0,04
≤ 8 ans	0,15	0,17	0,04

Supposons que nous ayons une variable de réponse binaire, Y , avec $Y = 1$ dénotant un cas et $Y = 0$ dénotant éventuelles, x . Nous supposons que la valeur de Y est connue pour chacune des N unités d'une population cible donnée, mais qu'au moins certaines composantes de x sont inconnues. Nous stratifions la population en cas et en témoins, tirons un échantillon dans chaque strate d'après les variables que nous connaissons pour toutes les unités, et mesurons les valeurs des covariables manquantes pour les unités échantillonnées (en pratique, l'échantillon de témoins est souvent tiré à partir de l'ensemble de la population, plutôt que parmi les unités pour lesquelles $Y = 0$. Si la

influence considérable sur le choix des méthodes d'échantillonnage adoptées pour réaliser ces études (voir Hartge, Brinton, Rosenthal, Cahill, Hoover et Waksberg 1984 qui donnent aussi une description de plusieurs études similaires), de sorte qu'il s'agit d'un point de départ naturel.

Exemple 1

En 1977-1978, le National Cancer Institute et l'Environmental Protection Agency des États-Unis ont réalisé une étude cas-témoins sur la population afin d'examiner les effets des rayonnements ultraviolets sur le cancer de la peau de type non-mélanome au cours d'une période d'un an (Hartge, Brinton, Rosenthal, Cahill, Hoover et Waksberg 1984; Fears et Gail 2000). L'étude a été réalisée dans huit régions géographiques où l'intensité des rayonnements solaires ultraviolets diffèrait. Dans chaque région, un échantillon de personnes atteintes d'un cancer de la peau de type non-mélanome âgées de 20 à 74 ans et un échantillon de témoins sélectionnés dans la population générale ont été interviewés par téléphone afin d'obtenir des renseignements sur les facteurs de risque. Pour chaque région, un échantillon aléatoire simple de 450 cas et un échantillon supplémentaire de 50 cas du groupe des 20 à 49 ans ont été sélectionnés en vue d'une prise de contact. Pour les témoins, 500 ménages ont été échantillonnés dans de Mitosky-Waksberg (Waksberg 1978). On s'est efforcé, dans la mesure du possible, d'interviewer tous les adultes de 65 à 74 ans, ainsi qu'une personne de chaque sexe de 20 à 64 ans choisie au hasard. En outre, un deuxième échantillon de Mitosky-Waksberg contenant de 500 à 2 100 ménages a été sélectionné et des renseignements ont été recueillis au sujet de tous les adultes de 65 à 74 ans. Cela a donné un échantillon d'environ 3 000 cas et un échantillon d'environ 8 000 témoins, le taux d'échantillonnage des cas étant environ 300 fois celui des témoins, selon l'âge.

Le deuxième exemple revêt une importance particulière pour moi, car il correspond à la première incursion que Chris Wild et moi-même avons faite dans ce domaine.

Exemple 2

L'étude de la méningite à Auckland a été exécutée à la demande du ministère de la Santé et du Conseil de recherche sur la santé de la Nouvelle-Zélande en vue d'examiner les facteurs de risque de méningite chez les jeunes enfants parmi lesquels la progression de la maladie prenait des proportions épidémiques (voir Baker, McNicholas, Garrett, Jones, Stewart, Kobenstein et Lennon 2000). L'étude avait pour population cible l'ensemble des enfants de moins de neuf ans dans la région d'Auckland entre 1997 et 2000. Tous les cas de méningite relevés dans le groupe d'âge cible au cours des trois années qu'a duré l'étude ont été inclus, ce qui a donné environ 250 cas. Un nombre

Études cas-témoins basées sur la population

Alastair Scott¹

Résumé

Nous discutons de méthodes d'analyse des études cas-témoins pour lesquelles les témoins sont sélectionnés selon un plan de sondage complexe. La méthode la plus simple est l'approche du sondage standard basée sur des versions pondérées des équations d'estimation pour la population. Nous examinons aussi des méthodes plus efficaces et comparons leur degré de robustesse aux erreurs de spécification du modèle dans des cas simples. Nous discutons également brièvement des études familiales cas-témoins, pour lesquelles la structure intragrappe présente un intérêt en soi.

Mots clés : Études cas-témoins; échantillonnage sélectif; échantillonnage rétrospectif; pondération.

1. Introduction

L'étude cas-témoins, dans laquelle des échantillons distincts sont tirés parmi les « cas » (disons, les personnes présentant une maladie d'intérêt) et parmi les « témoins » (personnes n'ayant pas la maladie), est l'une des méthodes d'étude les plus fréquentes en recherche sur la santé. En fait, Breslow (1996) a qualifié ce genre d'étude de « pivot de l'épidémiologie ». Nous nous concentrerons ici sur les applications biostatistiques, mais le plan de base représente une stratégie d'échantillonnage efficace dans les situations où les cas sont rares et est aussi utilisé fréquemment dans de nombreux autres domaines (commerce, sciences sociales, écologie, études de marché, par exemple). On a notamment assisté dans la littérature économique au développement parallèle de la plupart de la théorie de l'échantillonnage basé sur les choix (voir, par exemple, Manski et McFadden 1981; Cosslett 1981).

Il existe deux types fondamentalement différents d'études cas-témoins, à savoir les études avec cas-témoins appariés, dans lesquelles chaque cas est apparié à un ou à plusieurs témoins, et les études avec cas-témoins non appariés, dans lesquelles les échantillons de cas et de témoins sont tirés indépendamment, quoiqu'il puisse exister un vague « appariement fréquentiste », l'échantillon de témoins étant réparti entre des strates définies par des variables démographiques de base de façon telle que la distribution de ces variables dans l'échantillon de témoins soit la même que celle prévue dans l'échantillon de cas. Nous ne nous intéressons ici qu'aux études sans appariement et, plus précisément, uniquement à la catégorie restreinte d'études sur la population dans lesquelles les témoins (et, à l'occasion, les cas également) sont sélectionnés en utilisant des méthodes standard d'échantillonnage.

Une excellente introduction à l'échantillonnage cas-témoins et à ses points forts et ses inconvénients éventuels est donnée dans Breslow (1996, 2004). L'un des plus grands

défis que doit relever toute personne qui conçoit une étude de ce genre consiste à s'assurer que les témoins soient réellement sélectionnés à partir de la même population, selon les mêmes protocoles, que les cas. Comme l'a dit Miettinen (1985), les cas et les témoins « devraient être représentatifs de la même expérience fondamentale » [traduction]. Le fait qu'au début, des mesures adéquates n'aient pas été prises à cet égard lors de certaines études a suscité des doutes au sujet de l'échantillonnage cas-témoins chez de nombreux chercheurs. Une discussion approfondie des principes qui devraient régir la sélection des témoins figure dans Wacholder, McLaughlin, Silverman et Mandel (1991). Puisque l'essence de l'échantillonnage tient aux méthodes suivies pour tirer des échantillons représentatifs à partir d'une population cible, il est devenu naturel de penser aux méthodes de sondage pour obtenir les témoins. De plus en plus fréquemment, au cours des quelque 25 dernières années, les témoins (et parfois les cas également) ont été sélectionnés en utilisant des plans de sondage complexes stratifiés à plusieurs degrés. Le lecteur trouvera au chapitre 9 de Korn et Graubard (1999) un bon historique de cette évolution.

L'analyse de ce genre d'études est un sujet tout indiqué pour le présent article, car Joe Waksberg lui-même a été l'un des principaux artisans de l'adoption des méthodes de sondage (et de la composition aléatoire, en particulier) pour l'obtention des témoins (voir, par exemple, Waksberg 1998, ainsi que Digiacanto et Waksberg 2002).

2. Exemples

Nous commençons par deux exemples en vue d'illustrer le genre de problèmes que nous voulons résoudre. Le premier est typique des études à grande échelle réalisées par le National Cancer Institute, aux employées duquel nous Joe Waksberg, et ses collègues de Westat ont eu une

Deville et Lavallée, pour leur part, présentent les fondements théoriques généraux de la méthode du partage des poids appliquée au sondage indirect. Ils définissent l'important concept d'une matrice de liens dans le sondage indirect, qui précise comment les éléments de la population échantillonnée sont liés à la population cible et qui pondère ces liens pour que l'estimation soit sans biais. Ils examinent d'importantes propriétés de la matrice de liens et définissent les conditions nécessaires et suffisantes à une matrice de liens optimale. Les auteurs démontrent leur théorie avec quelques exemples intéressants.

Deville et Mauney-Bertrand étudient la détermination d'un plan de sondage et d'une méthode d'estimation dans le cas d'une enquête sur la fréquentation touristique. Le problème principal posé par ce type d'enquêtes est l'absence d'une base de sondage permettant d'atteindre directement les touristes. Pour contourner ce problème, les auteurs suggèrent d'échantillonner des services destinés principalement aux touristes. On se trouve donc dans une situation de sondage indirect pour laquelle la méthode généralisée du partage des poids est utilisée pour obtenir les estimations des paramètres d'intérêt. Certaines extensions à la méthode sont nécessaires. Les auteurs portent leur attention plus particulièrement à l'une d'elles et la décrivent en détail.

Félix-Medina et Monjardín examinent une variante de l'échantillonnage par dépiage de liens. Pour ce faire, ils ont recours à une approche Bayésienne pour construire des estimateurs de taille de population. Cependant, pour que les inférences sur la taille de la population soient robustes aux spécifications erronées du modèle hypothétique, les auteurs font ces inférences selon une approche fréquentiste basée sur le plan. L'étude de simulation montre que les estimateurs proposés donnent de meilleurs résultats que les estimateurs du maximum de vraisemblance qui sont utilisés actuellement.

Dans leur article, Dorfman, Lent, Leaver et Wegman comparent la méthodologie de l'indice des prix à la consommation du Royaume-Uni et celle des États-Unis en s'appuyant sur les mêmes données scanographiques. Ils concluent que pour la population étudiée, l'approche du Royaume-Uni, qui suppose une stratification plus rigoureuse et, ce qui importe davantage, un échantillonnage intrastate par choix raisonné plus restrictif que l'approche par échantillonnage probabiliste des États-Unis, donne un meilleur indice superlatif estimé cible. C'est le cas quel que soit l'estimateur (ratio des moyennes, moyenne géométrique ou moyenne des ratios) utilisé pour produire l'indice de prix de bas niveau.

Dans leur article, Thomas, Raghunathan, Schenker, Katzoff et Johnson ont recours à l'imputation multiple pour analyser des données qui, à cause d'un plan de sondage matriciel, présentent des valeurs manquantes. Dans le plan de sondage matriciel, seul un sous-ensemble de questions sont posées à chaque répondant afin de réduire le fardeau de réponse. Les auteurs ont élaboré une méthode permettant de créer des formulaires de sondage matriciel. Chaque formulaire contient un sous-ensemble de questions à poser à des répondants choisis au hasard. La méthode est conçue pour que chaque formulaire comprenne des questions qui permettent de prédire les réponses aux questions exclues, ce qui permet de déduire une partie de l'information exclue. La méthode proposée et l'imputation multiple sont évaluées à l'aide de données tirées de la National Health and Nutrition Examination Survey.

Harold Mantel, Rédacteur en chef délégué

Dans ce numéro

Ce numéro de *Techniques d'enquête* débute par le sixième article de la série d'articles annuels sollicités dédiée à Joseph Waksberg. C'est avec tristesse que nous soulignons sa disparition survenue en janvier 2005. Une brève biographie de Joseph Waksberg a été publiée dans notre numéro de juin 2001, qui présentait aussi le premier article de la série. Pour plus d'informations sur la vie et l'oeuvre de Joseph Waksberg, voir l'article de Statistical Science (Vol. 15, No 3) "A Conversation with Joseph Waksberg" par David Morganstein et David Markter disponible à <http://projecteuclid.org/Dienst/UI/1.0/Home>. Gordon Brackstone, Sharon Lohr et Wayne Fuller – d'avoir choisi Alastair Scott comme auteur de l'article du prix Waksberg de cette année.

Dans son article intitulé « Etudes cas-témoins basées sur la population », Scott traite des études cas-témoins pour lesquelles les témoins sont obtenus à partir d'une enquête par sondage complexe. En utilisant l'exemple de la régression logistique, il démontre que les estimations pondérées de l'enquête peuvent être passablement inefficaces en raison du poids assez faible attribué aux cas. Établissant une analogie avec l'estimation du maximum de vraisemblance, il propose ensuite une solution de rechange simple et bien plus efficace, mais biaisée pour l'origine. À l'aide d'exemples, il démontre les propriétés d'efficacité et de robustesse. Finalement, il examine brièvement la question des études cas-témoins portant sur des familles.

Pour sa part, Kott examine la pondération par calage pour corriger les erreurs de non-réponse et de couverture. Il donne une description générale de l'estimation par calage et clarifie l'approche de la forme fonctionnelle d'Estayo et de Särndal à la fonction générale de calage. Il analyse ensuite les propriétés de cette méthode de calage pour corriger les erreurs dues à la non-réponse des unités et les erreurs de couverture en fonction d'un modèle de quasi-randomisation. Il conclut avec un exemple empirique et une analyse de certaines questions importantes.

À l'aide d'une étude de simulation, Reiter, Raghunathan et Kinney étudient les conséquences de ne pas tenir compte des variables du plan de sondage lors de l'élaboration de modèles d'imputation dans un contexte d'imputation multiple. Ils montrent que les biais potentiels peuvent être réduits en incluant ces variables dans le modèle d'imputation au moyen d'un modèle à effets fixes ou à effets mixtes. Les auteurs concluent que les imputeurs auraient intérêt à inclure comme prédicteurs toutes les variables liées aux variables imputées, en particulier les variables du plan de sondage, de manière à satisfaire aux conditions de l'hypothèse habituelle selon laquelle la non-réponse est ignorable.

Dans leur article, Funaoka, Saigo, Sitter et Tolda se penchent sur l'utilisation des estimateurs de variance bootstrap dans l'échantillonnage stratifié à plusieurs degrés lorsque les fractions de sondage sont grandes. Ils proposent une méthode bootstrap de type Bernoulli qui donne des estimations cohérentes de la variance bootstrap lorsque l'échantillonnage aléatoire simple sans remise est utilisé à chaque degré. La méthode proposée est simple à appliquer et peut être étendue à n'importe quel nombre de degrés d'échantillonnage sans trop de complications. La méthode est illustrée à l'aide d'une étude de simulation limitée qui utilise des données de l'Enquête nationale sur les prix menée en 1997 au Japon.

Dans leur article, Kozak et Verma comparent l'approche géométrique de stratification, proposée par Gunning et Horgan (2004), avec deux approches d'optimisation : l'algorithme de Lavallée-Hidiroglou (Lavallée et Hidiroglou 1988) et l'algorithme d'optimisation proposé par Kozak (2004). Au moyen de cinq populations fictives de différentes tailles, les trois méthodes sont comparées selon deux scénarios : une comparaison du coefficient de variation (c.v.) résultant selon une taille d'échantillon fixe et une comparaison des tailles d'échantillon résultant selon un niveau de précision fixe.



The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 32, numéro 2, décembre 2006

Table des matières

Dans ce numéro.....	135
Article Sollicite Waksberg	
Alastair Scott	
Études cas-témoins basées sur la population.....	137
Regular Papers	
Phillip S. Kott	
Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture.....	149
Jerome P. Reiter, Trivellore E. Raghunathan et Satkartar K. Kinney	
L'importance de la modélisation du plan d'échantillonnage dans l'imputation multiple pour les données manquantes.....	161
Fumio Funakoka, Hiroshi Saigo, Randy R. Sitter et Tsutom Tolda	
Bootstrap de type Bernoulli pour l'échantillonnage stratifié à plusieurs degrés.....	169
Marcin Kozak et Med Ram Verma	
Approche de la stratification par une méthode géométrique et par optimisation : Une comparaison de l'efficacité.....	177
Jean-Claude Deville et Pierre Lavallée	
Sondage indirect : Les fondements de la méthode généralisée du partage des poids.....	185
Jean-Claude Deville et Myriam Maunay-Bertrand	
Extensions de la méthode d'échantillonnage indirect et son application aux enquêtes dans le tourisme.....	197
Martin H. Félix-Medina et Pedro E. Monjardin	
Combinaison de l'échantillonnage par dépiçage de liens et de l'échantillonnage en grappes pour estimer la taille de populations cachées : Une approche assistée par la méthode bayésienne.....	207
Alan H. Dorfman, Janice Lent, Sylvia G. Leaver et Edward Wegman	
Plans de sondage pour les indices des prix à la consommation.....	217
Neal Thomas, Trivellore E. Raghunathan, Nathaniel Schenker, Myron J. Katzoff et Clifford L. Johnson	
Une évaluation des méthodes d'échantillonnage matriciel à l'aide de données provenant de la National Health and Nutrition Examination Survey.....	241
Remerciements.....	259

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertorié dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

D. Royce

Anciens présidents

G.J. Brackstone

R. Platek

J. Kovar

H. Mantel

E. Rancourt

M.P. Singh

Ancien rédacteur en chef

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

M. Thompson, *University of Waterloo*

Y. Tillé, *Université de Neuchâtel*

R. Valliant, *JPSM, University of Michigan*

V.J. Verna, *Università degli Studi di Siena*

K.M. Wolter, *Iowa State University*

C. Wu, *University of Waterloo*

A. Zaslavsky, *Harvard University*

T.J. Rao, *Indian Statistical Institute*

J.N.K. Rao, *Carleton University*

N.G.N. Prasad, *University of Alberta*

D. Pfeffermann, *Hebrew University*

G. Nathan, *Hebrew University*

P. Lavallée, *Statistique Canada*

P. Lahiri, *JPSM, University of Maryland*

R. Kott, *National Agricultural Statistics Service*

D. Judkins, *Westat Inc.*

M.A. Hidroglou, *Office for National Statistics*

J. Gambino, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.L. Eitinge, *U.S. Bureau of Labor Statistics*

P. Cantwell, *U.S. Bureau of the Census*

J.M. Brick, *Westat Inc.*

D.A. Binder, *Statistique Canada*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

REDACTEURS ADJOINTS

J.-F. Beaumont, P. Dick, D. Haziza, Z. Patak, S. Rubin-Bleuer et W. Yung, *Statistique Canada*

T.J. Rao, *Indian Statistical Institute*

J.N.K. Rao, *Carleton University*

N.G.N. Prasad, *University of Alberta*

D. Pfeffermann, *Hebrew University*

G. Nathan, *Hebrew University*

P. Lavallée, *Statistique Canada*

P. Lahiri, *JPSM, University of Maryland*

R. Kott, *National Agricultural Statistics Service*

D. Judkins, *Westat Inc.*

M.A. Hidroglou, *Office for National Statistics*

J. Gambino, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.L. Eitinge, *U.S. Bureau of Labor Statistics*

P. Cantwell, *U.S. Bureau of the Census*

J.M. Brick, *Westat Inc.*

D.A. Binder, *Statistique Canada*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

D. Steel, *University of Wollongong*

L. Stokes, *Southern Methodist University*

J. Kovar, *Statistique Canada*

H. Mantel, *Statistique Canada*

E. Rancourt

M.P. Singh

J. Reiter, *Duke University*

L.-P. Rivest, *Université Laval*

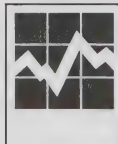
N. Schenker, *National Center for Health Statistics*

F.J. Scheuren, *National Opinion Research Center*

C.J. Skinner, *University of Southampton*

E. Stasny, *Ohio State University*

Techniques d'enquête



7748

Une revue éditée par Statistique Canada

Décembre 2006 • Volume 32 • Numéro 2

Publication autorisée par le ministre
responsable de Statistique Canada

© Ministre de l'Industrie, 2006

Tous droits réservés. Le produit ne peut
être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence.

Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnelle, d'organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication de résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada, K1A 0T6.

Décembre 2006

N° 12-001-XPB au catalogue

Périodicité : semestrielle

ISSN 0714-0045

Ottawa





Numéro 2

•

Volume 32

•

Décembre 2006

Une revue
éditée
par Statistique Canada

N° 12-001-XPB au catalogue

Techniques d'enquête

